Machine Learning for Marketing Decision Support

Doctoral Thesis

to acquire the academic degree of
doctor rerum politicarum
(Doctor of Economics and Management Science)

submitted to

School of Business and Economics
Humboldt-Universität zu Berlin

by

M.Sc. Johannes Sebastian Haupt

President of Humboldt-Universität zu Berlin:
Prof. Dr.-Ing. Dr. Sabine Kunst

Dean of the School of Business and Economics:
Prof. Dr. Daniel Klapper

Reviewers:
1. Prof. Dr. Stefan Lessmann
2. Prof. Dr. Daniel Guhl

Date of Colloquium: 24 June 2020

Abstract

The digitization of the economy has fundamentally changed the way in which companies interact with customers and made customer targeting a key intersection of marketing and information systems. Marketers can choose very specifically which customers to serve with a marketing message based on detailed demographic and behavioral information. Building models of customer behavior at the scale of modern customer data requires development of tools at the intersection of data management and statistical knowledge discovery. The application of these models for successful targeting requires deep understanding of the underlying marketing decision problem and awareness of the ethical implications of data collection.

This dissertation widens the scope of research on predictive modeling by focusing on the intersections of model building with data collection and decision support. Its goals are 1) to develop and validate new machine learning methods explicitly designed to optimize customer targeting decisions in direct marketing and customer retention management and 2) to study the implications of data collection for customer targeting from the perspective of the company and its customers.

The thesis addresses the first goal by proposing methods that utilize the richness of e-commerce data, reduce the cost of data collection through efficient experiment design and address the targeting decision setting during model building. The underlying state-of-the-art machine learning models scale to high-dimensional customer data and can be conveniently applied and adapted by practitioners. These models further address the problem of causal inference that arises when the causal attribution of customer behavior to a marketing incentive is difficult. Marketers can directly apply the model estimates to identify profitable targeting policies in applications with complex cost structures.

The analyses addressing the second goal of the thesis quantify the savings potential of efficient experiment design and the monetary cost of an internal principle of data privacy. Practitioners can follow the proposed methodology to evaluate internally collected data like a commodity and make informed decisions. An analysis of data collection practices in direct marketing emails reveals the ubiquity of tracking mechanisms without user consent in e-commerce communication. These results form the basis for a machine-learning-based system for the detection and deletion of tracking elements from emails.

Keywords: Customer Targeting, Machine Learning, Decision Support, Data Privacy

Zusammenfassung

Die Digitalisierung der Wirtschaft hat die Interaktion zwischen Firmen und Kunden grundlegend verändert und macht das Customer Targeting zu einer wichtigen Schnittmenge von Marketing und Wirtschaftsinformatik. Marketingtreibende können auf Basis von soziodemografischen und Verhaltensdaten gezielt einzelne Kunden mit personalisierten Botschaften ansprechen. Die Erstellung von Modellen zur Vorhersage von Kundenverhalten, die hochdimensionalen, modernen Kundendaten gerecht werden, erfordert die Weiterentwicklung von Methoden an der Schnittstelle von Datenmanagement und statistischer Analyse. Die Anwendung dieser Modelle für das gewinnbringende Auswahl individueller Zielkunden erfordert umfassendes Verständnis der zugrunde liegenden Entscheidungsprobleme im Marketing und ein Bewusstsein für die ethischen Aspekte der Datenerfassung.

Die vorliegende Arbeit erweitert die Perspektive der Forschung im Bereich der modellbasierten Vorhersage von Kundenverhalten durch ihren Fokus auf die Schnittstellen der statistischen Modellierung zur Datenerfassung und Entscheidungsunterstützung. Ziel der Arbeit ist 1) die Entwicklung und Validierung neuer Methoden des maschinellen Lernens, die explizit darauf ausgelegt sind, die Profitabilität des Customer Targeting im Direktmarketing und im Kundenbindungsmanagement zu optimieren, und 2) die Untersuchung der Datenerfassung mit Ziel des Customer Targeting aus Unternehmens- und Kundensicht.
Die Arbeit adressiert das erste Ziel durch die Entwicklung von Methoden, welche den Umfang von E-Commerce-Daten nutzbar machen und die Rahmenbindungen der Marketingentscheidung während der Modellbildung berücksichtigen. Die zugrundeliegenden Modelle des maschinellen Lernens skalieren auf hochdimensionale Kundendaten und ermöglichen die unkomplizierte Anwendung und Erweiterung in der Praxis. Die vorgeschlagenen Methoden basieren zudem auf dem Verständnis des Customer Targeting als einem Problem der Identifikation von Kausalzusammenhängen. Die Modellschätzung sind für die Umsetzung profitoptimierter Zielkampagnen unter Berücksichtigung komplexer Kostenstrukturen in der Praxisanwendung ausgelegt.
Die Arbeit adressiert das zweite Ziel durch die Quantifizierung des Einsparpotenzials effizienter Versuchsplanung bei der Datensammlung und der monetären Kosten der Umsetzung des Prinzips der Datensparsamkeit. Die vorgeschlagene Methodik erlaubt Praxisanwendern die Evaluation potentieller Daten als Produktionsfaktor zur Modellschätzung, um auf dieser Basis fundierte Entscheidungen zu deren Erhebung treffen zu können. Eine Analyse der Datensammlungspraktiken im E-Mail-Direktmarketing zeigt zudem, dass eine Überwachung des Leseverhaltens in der Marketingkommunikation von E-Commerce-Unternehmen ohne explizite Kundenzustimmung weit verbreitet ist. Diese Erkenntnis bildet die Grundlage für ein auf maschinellem Lernen basierendes System zur Erkennung und Löschung von Tracking-Elementen in E-Mails.

Schlagworte: Direktmarketing, Maschinelles Lernen, Entscheidungsunterstützung, Datenschutz

Acknowledgments

I wish to express my deepest gratitude to my supervisor, Prof. Stefan Lessmann, whose excellent teaching inspired me to start a PhD in machine learning. His support has given me the chance to pursue ideas that still fascinate me and his input and guidance have given me the means to turn these ideas into research. I want to express my gratitude to my second supervisor, Prof. Daniel Guhl, and to Prof. Dr. Daniel Klapper, who have introduced me to the field of quantitative marketing. I also thank Prof. Dr. Bart Baesens for inviting me to work with his group at KU Leuven.

I am indebted to my coauthors, especially Prof. Ben Fabian, Dr. Annika Baumann, Benedict Bender, Daniel Jacob, Robin Gubela and Fabian Gebert, who have gifted me their knowledge and time on countless occasions. I am grateful to my colleagues and my fellow PhD students, among them Dr. Sebastian Gabel, Dr. Alona Zharova, Nikita Kozodoi, Narine Yegoryan, Tobias König, Marius Sterling, Elizaveta Zinovyeva, Alisa Kim, Gary Mena, Elias Baumann, Eugen Stripling and many others, for sharing the ups and downs of this path. Thank you all for the exciting discussions and happy lunches that we shared over the years.

I would like to thank the students of the faculty for their curiosity and hard work. I would also like to thank Anna-Lena Bujarek and the Humboldt Lab for Empirical and Quantitative Research for their support.

My deepest thanks go out to my parents, Werner and Evelyn, and my sister Anja for a lifetime of care and to my wonderful wife Anlin. This thesis would not exist without her encouragement, support and patience to discuss statistics in her free time.

Contents

List of Figures

List of Tables

Abbreviations

ANN	Artificial Neural Network
ATE	Average Treatment Effect
AUC	Area under the Receiver-Operating-Characteristic-Curve
AUC-PR	Area under the Precision-Recall-Curve
BBM	Best-Base Model
CATE	Conditional Average Treatment Effect
CF	Causal Forest
CRM	Customer Relationship Management
KDD	Knowledge Discovery in Databases
CLV	Customer Lifetime Value
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CSS	Cascading Style Sheets
CVT	Class Variable Transformation
DR	Doubly-Robust Estimator
DSS	Decision Support System
DT	Decision Tree
ERT	Extremely Randomized Trees
FNN	Feed-Forward Neural Network
GA	Genetic Algorithm
GER	Germany
GLM	Generalized Linear Model
IP	Internet Protocol
IPW	Inverse Probability Weighting
ITE	Individualized Treatment Effect
ITM	Interaction Term Method
kNN	k-Nearest Neighbors
LWUM	Lai's Weighted Uplift Method
MAE	Mean-Absolute Error
MSE	Mean-Squared Error
MTA	Mail Transfer Agent
MUA	Mail User Agent
NLL	Negative Log-Likelihood
PAES	Profit-Agnostic Ensemble Selection
PCES	Profit-Conscious Ensemble Selection
PII	Personally Identifiable Information
RCT	Randomized Controlled Trial
RF	Random Forest
RFM	Recency, Frequency and Monetary Value
RMSE	Root-Mean-Squared Error
ROC	Receiver-Operating-Characteristic-Curve
SD	Standard Deviation
SMOTE	Synthetic Minority Oversampling Technique
SMTP	Simple Mail Transfer Protocol
SVM	Support Vector Machine
TCIA	Treatment-Covariate Interaction Method
TDL	Top-Decile Lift
TOL	Transformed Outcome Loss
TS	Theil-Sen Regression
UK	United Kingdom of Great Britain
UML	Unified Modeling Language
UMPPM	Uplift Modeling Process for Marketing
URL	Uniform Resource Locator
US	United States of America
VIF	Variable Inflation Factor

Chapter 1

Introduction

Customer targeting is a key intersection of marketing and information systems. At the core of this development lies the digitization of the economy that has fundamentally changed the way in which companies interact with customers. Marketers can choose very specifically which customers to serve with a marketing message and personalize the message based on detailed demographic and behavioral information. Direct mail campaigns have increasingly substituted email in place of traditional print mail, which provides a cost-efficient digital channel (Hartemo, 2016) and is customizable at large scale (Sahni et al., 2018). Traditional print advertising has expanded from billboards and magazines to the internet, where marketers bid in real-time to place their message in the available advertising space for specific customers (Stange & Funk, 2014). Smartphone applications and push notifications allow companies to reach out to customers directly and instantaneously with messages that are optimized for customers' habits, locations and current activities (Dubé et al., 2017; Lian et al., 2019). These opportunities for customer targeting are made possible by developments in the infrastructure to collect customer data, process it on a large scale and automatize the targeting decision (Ansari & Mela, 2003).

The technological development of digital marketing and the underlying infrastructure has impacted competition between companies. The organizational and technological requirements of digitization allow companies to compete through the optimization of operational decision-making (Hormozi & Giles, 2004) and target marketing (Yang et al., 2014). Within this competition, data access has become a competitive advantage and an issue of corporate social responsibility (Pollach, 2011). The feasibility of individual discounts is a form of personalized pricing that, on one hand, may increase price discrimination towards customers (Acquisti & Varian, 2005) and, on the other hand, may strengthen overall price competition in the market (Shaffer & Zhang, 2002). The net impact of data collection and processing on individual and societal welfare depends strongly on the value of the application or industry and the associated risks to consumers (Acquisti et al., 2016). Within customer targeting, technological competition has lead to the adaption of the statistical and computational tools required for the large-scale aggregation and processing of personal data and automated decision making.

Processing information to build models of customer behavior at the scale of modern customer data has required continuous development of tools at the intersection of data management, statistical knowledge discovery and marketing domain knowledge (Shaw et al., 2001). Generative models paired with Bayesian inference methods remain popular in the marketing literature (Rossi et al., 2005; Ruiz et al., 2017), but flexible model specifications are difficult to scale to more than hundreds of unique customers and products or tens of thousands of observations (Ishigaki et al., 2018; Jacobs et al., 2016). The information systems literature was an

early adopter of machine learning models for the prediction of customer behavior from high-dimensional datasets (Agrawal et al., 1993; Bose & Xi, 2009). Machine-learning models have been successfully applied to model customer choice for hundreds of thousands of customers (Gabel, 2019), hundreds of thousands of products (Grbovic et al., 2015), and high-cardinal variables typical to socio-demographic data (Moeyersoms & Martens, 2015). Recent advances have expanded the applicability of these models to complex data structures. This allows behavior modeling for large-scale panel data (Chen et al., 2015; Martens et al., 2016) and a more effective use of text data (De Caigny et al., 2019) and network graphs (Backiel et al., 2016). The progress of modeling customer behavior in computation and statistics is coupled with research on the utilization of these models for profitable marketing decision making and research on the collection of the customer data that serves as input to the models.

The successful application of machine-learning models for customer targeting requires a deep understanding of the underlying decision problem. Research on information systems acknowledges this decision support component and aims to reconcile it with classification models through the paradigm of cost-sensitive learning (Elkan, 2001). Kim and Moon (2012) and Verbeke et al. (2012) propose methods that address the uncertainty of the future value of customers to the company when evaluating churn models. Glady et al. (2009) and Kim et al. (2013) integrate the estimated customer value and costs of different measures into the targeting model. The marketing literature approaches customer targeting with a stronger focus on the decision problem and assumptions underlying specific applications. The definition and estimation of the profit generated by a customer for the company is a difficult problem that is addressed by extensive research on customer lifetime value (e.g. Chan et al., 2011; Kumar et al., 2008). For example, given the close connection between a customers' product usage and their decision to remain customers, Ascarza and Hardie (2013) propose to model the customer retention decision and customer value jointly. Based on the value of the customer and the expected effect and cost of the marketing action, Hansotia and Rukstales (2002) provide an analysis of campaign targeting as a decision problem. Hitsch and Misra (2018) apply their proposed policy to optimize how many and which customers to target in a print campaign. Despite these advances, the combination of scalable models of customer behavior with profitable decision making remains a challenging area for research due to the diversity in decision settings of the applications in which customer targeting is applied.

The collection of personal data used as model input introduces an ethical dimension to the application of customer targeting. The expanding collection of customer information has raised concerns of stakeholders over compliance and the security of this information and concerns of customers for their privacy (Anderson & Moore, 2006). Privacy concerns have been shown to lead to a loss in customer trust towards companies and undermine the effectiveness of personalized marketing (Goldfarb & Tucker, 2011), despite an imperfect match between customers' self-reported preference for privacy and their observed behavior (Nofer et al., 2014). The growing awareness for privacy is contrasted in practice by the ubiquity of data acquisition from information brokers or collection without explicit consent, for example in the form of browsing data or email reading behavior. The contemporary spread of third-party tracking allows data

brokers to track customer behavior beyond a single brick-and-mortar store or a chain of stores to detailed movement within the ecosystem of web properties and online shops (Bucklin & Sismeiro, 2009; Mayer & Mitchell, 2012). Data brokers further enrich tracking data with data from other sources, e.g. social media profiles (Bradlow et al., 2017) and email communication (Grbovic et al., 2015). More recently, the ubiquity of smartphones has extended the collection of behavioral data back to the offline world by providing the means to continuously collect location data and inferred customer activities (Dubé et al., 2017). Increasing awareness among customers and a stronger regularization on data collection through policies like the European General Data Protection Regulation[1] aim to align business goals with customer interests. The extent to which data is collected or acquired for customer targeting should be seen as a strategic management decision that relates to business ethics (Hand, 2018) and customer trust (Bansal et al., 2015), which strengthens the need for research on data privacy in the context of customer targeting (Goldfarb & Tucker, 2011)

Within the context of decision-making, the modeling of customer behavior is only one of several key steps in the process of customer targeting, which is preceded by data collection and utilized as part of a profit-maximizing targeting policy. The motivation of this thesis is to widen the scope of research on predictive modeling by focusing on the intersections of model building with data collection and decision support. Its goals are 1) to develop and validate new machine learning methods explicitly designed to optimize customer targeting decisions in direct marketing and customer retention management and 2) to study the implications of data collection for customer targeting from the perspective of the company and its customers.
The thesis addresses the first goal by proposing methods that utilize the richness of e-commerce data, reduce the cost of data collection through efficient experiment design and address the targeting decision setting during model building. The underlying state-of-the-art machine learning models scale to high-dimensional customer data and can be conveniently applied and adapted by practitioners. These models further address the problem of causal inference that arises when the causal attribution of customer behavior to a marketing incentive is difficult. Marketers can directly apply the model estimates to identify profitable targeting policies in applications with complex cost structures.
Collecting the data required to apply these methods and model customer behavior is a management decision. The analyses addressing the second goal of the thesis quantify the savings potential of efficient experiment design and the monetary cost of an internal principle of data reduction and data economy. Practitioners can follow the proposed methodology to evaluate internally collected data like a commodity and make informed decisions. An analysis of data collection practices in direct marketing emails reveals the ubiquity of tracking mechanisms without user consent in e-commerce communication. These results form the basis for a machine-learning-based system for the detection and deletion of tracking elements from emails.

The thesis addresses its goals through the ten essays summarized in Figure 1.1. The first five

[1]Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) available at http://data.europa.eu/eli/reg/2016/679/2016-05-04

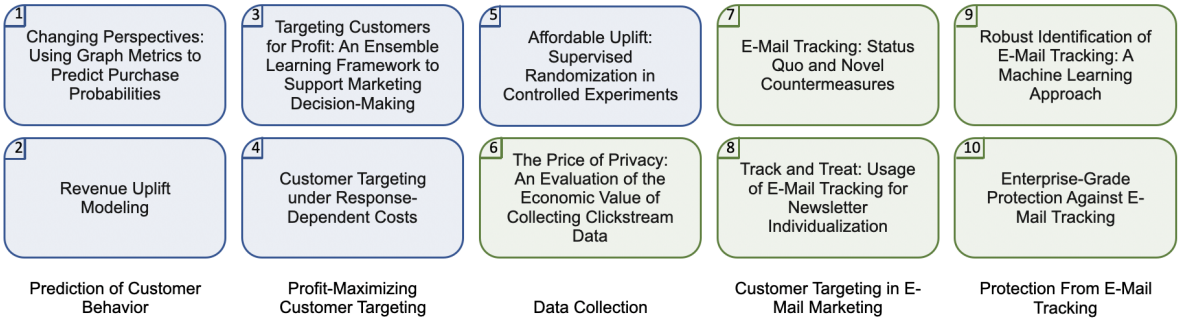


Figure 1.1: Structure of the ten essays comprising the thesis. Chapters 2–6 (blue) address customer target decision support for profit-maximization, Chapters 7–11 (green) the prerequisite data collection and customer privacy marked.

essays propose advances to profit-centered customer targeting, starting with the improvement of existing predictive models of customer response and the integration of the profit-maximization goal into model building, before addressing the targeting decision as a problem of causal inference and, finally, developing a method for efficient data collection.

Chapter 2 proposes an efficient way to aggregate customer behavior on the website over time using graph theory. Information on customer search behavior is an important input to customer purchase models and has been used extensively for customer segmentation in the form of aggregate measures, e.g. recency, frequency and monetary value (RFM) (Bauer, 1988). Aggregation of customer behavior into a set of variables reduces the complexity of the modeling problem at the cost of information. Recent research has thus emphasized the importance of expanding the traditional RFM measures to capture the pattern of individual customer interactions in more detail (Zhang et al., 2015). The aggregation of customer behavior data is particularly interesting for online data that is collected at a granular level in high quantity (Kohavi et al., 2004). Customer interactions in the e-commerce setting are collected at the level of a single shopping occasion in the form of clickstream data, where each webpage is logged as an interaction. The large number of unique webpages and the exponentially increasing combination of customer paths through the website makes it difficult to model the customer journey explicitly (Park & Park, 2016). This essay instead presents the clickstream of each visit as an expanding graph and studies the use of graph measures to reduce the clickstream to interpretable aggregate measures. These aggregate measures serve as single indicators of customer behavior that are predictive of purchase intention. The essay evaluates a range of graph measures for conversion prediction based on real-world data provided by two e-commerce shop. The empirical analysis confirms that the proposed measures of customer behavior have predictive value as measured by the precision-recall curve and top-decile lift. An interpretation is provided for the most valuable measures, which are identified to be closeness vitality and density followed by the graph radius and the number of self-loops and circles. The study illustrates that clickstream data is a relevant source of information that should be included in the statistical modeling of purchase intent. The identified aggregate measures show that graph measures can be a useful tool to identify behavior patterns and support that recurring search patterns may be indicative of purchase-oriented search rather than exploration behavior.

A profitable marketing campaign involves the decision which and how many individuals to target given the value of a potential customer and a cost to target her. Both the value and the cost of targeting may vary between customers. Cost-sensitive models integrate the potential decision outcomes and costs into the model building process to improve model precision for high-value customers or estimate the expected value of targeting a customer directly. Popular methods include the creation of synthetic observations for sparsely populated but profitable areas of the support (e.g. Zhu et al., 2017), integrating customer value into the loss function (Bahnsen et al., 2015b; Glady et al., 2009) or selecting model parameters with numerical optimization methods (Stripling et al., 2018). Optimization of an application-specific loss function during model estimation has been attempted for many common models, including regression (Finlay, 2010), neural networks (Kim et al., 2005) and decision trees (Bahnsen et al., 2015a). Alternatively, the cost-benefit setting of applications can be in-build into application-specific loss function (Lemmens & Gupta, 2017; Van den Poel & Prinzie, 2005). However, these optimization solutions require adaption of the loss function to the target domain and may impose restrictions on the form of the loss function that complicate model estimation.

Chapter 3 proposes a method to integrate established principles of statistical inference with marketing objectives in customer targeting. The proposed method uses ensemble selection on the level of estimates from a collection of statistical models to optimize an arbitrary objective function. The ensemble selection is based on greedy step-wise optimization (Caruana et al., 2004) that allows optimization of loss functions that are non-differentiable. The process of combining model estimates resembles how managers contextualize recommendation from model-based decision support systems (Fuller & Dennis, 2009) but shifts model post-processing into the modeling process. We hypothesize that a contextualization of the model development process improves the quality of targeting decisions. The essay evaluates profit-based ensemble selection against alternative models on twenty-five real-world marketing data sets from different industries. The alternative models consist of machine learning classifiers optimized on statistical loss functions and a linear model whose coefficients are optimized through numeric optimization on the loss function directly. The evaluation shows that the marginal campaign profit of profit-based ensemble selection is consistently higher the marginal profit derived from single models or ensemble selection based on a statistical loss. These results provide a clear and managerially meaningful measure of the business value of the extent to which the ensembling of models and the proposed model selection based on profit improve the decision quality.

Chapters 4 to 6 revisit the customer targeting problem as a problem of causal inference. Traditional customer response modeling implicitly relies on the assumption that receiving marketing treatment is a necessary condition for the intended customer action. When this assumption is fulfilled, e.g. for cold acquisition, the observed customer action after receiving treatment can be attributed to the treatment. Most customer targeting decisions are made in settings where customers have a natural propensity to behave in the intended way. For example, customers likely renew their telephone contract even in the absence of a marketing incentive. When the intended customer action occurs naturally, then a profitable treatment must show a positive incremental effect on the outcome. Ignoring customer behavior in the absence of treatment is

likely to inflate estimates of the profitability of the marketing treatment (Blake et al., 2015). Continuing the example, a campaign with no effect on customer retention will nevertheless exhibit a natural retention rate. Applications of customer targeting in settings with a natural propensity towards the outcome include economically important use cases like customer retention and direct marketing. The divergence between outcome prediction and treatment effect estimation as the basis for targeting decisions is particularly strong in coupon targeting, where the discount of implies an additional cost, and customer retention, where the chance and cost of free-riding is high (Ascarza, 2018). To estimate the incremental effect of marketing treatments, causal inference has reemerged as an important topic in customer targeting. This thesis expands on the literature on the estimation of the conditional average treatment effect as the basis of individual-level targeting decisions (Lo, 2002). In the setting of e-commerce, the number of customers and the technical ease of randomizing treatment assignment facilitate large-scale randomized controlled trials. To use this data to its full extent, this thesis focuses on estimating the treatment effect conditional on observed customer characteristics using machine learning models (Devriendt et al., 2018; Rzepakowski & Jaroszewicz, 2012).

Chapter 4 proposes an efficient method to model the conditional average treatment effect on revenue rather than conversion. The effect of marketing on a customer can be modeled as an effect on the customer's purchase incidence or their purchase value. Previous research has focused on the effect on purchase incidence (Hansotia & Rukstales, 2002; Rzepakowski & Jaroszewicz, 2012). However, modeling purchase incidence disregards any heterogeneity in individual spending. Modeling the treatment effect on revenue is a better strategy to optimize campaign profit (Hitsch & Misra, 2018), but an exact estimate of incremental customer value is not required for customer targeting in all practical settings. When the targeting policy is defined by a budget, e.g. targeting 10% of prospects, the targeting decision requires only an ordering of customers. In these settings, the model efficiency can be increased by binarization of the continuous target variable into profitable and unprofitable prospects (Bodapati & Gupta, 2004). The essay proposes a method that combines the discretization of customer revenue with causal modeling. The model estimates a transformed outcome variable that combines the treatment-based transformation of Tian et al. (2014) with the binarization of the customer value. This method avoids the complexity of causal inference techniques that require the estimation of more than one model and returns a profitability score for each customer. The proposed revenue transformation is computationally simple and flexibly accommodates any model specification, including standard machine learning algorithms. The study uses real-world data provided by an e-commerce shop to evaluate the revenue transformation against alternative transformation approaches and models of purchase propensity. The proposed method increases incremental revenue while introducing little additional complexity to model estimation.

Chapter 5 provides a generalization of analytical targeting policies to settings with costs that depend on the customer response. The common analysis of the customer targeting decision assumes that a fraction of customers with the strongest response to treatment is targeted (Devriendt et al., 2018) or that a customer is targeted if the estimated effect of the treatment is higher than the cost of applying the treatment (Hansotia & Rukstales, 2002). However, many

applications in direct marketing include costs that are uncertain at the time of the targeting decision because they are realized only when the customer accepts the marketing offer. These response-dependent costs are present whenever a marketing incentive is conditional on a profitable customer action. Companies use conditional incentives regularly in the form of discounts and the most salient applications have attracted much research, e.g. customer retention (Ascarza & Hardie, 2013; Backiel et al., 2016) or coupon targeting (Gubela et al., 2019; Sahni et al., 2016). Because the treatment cost is conditional on the customer action, the uncertainty about the customer action translates into uncertainty about the realization of the cost of the incentive. The essay provides a comprehensive analysis of the coupon targeting decision under response-dependent costs and proposes a model specification to efficiently estimate the necessary decision variables. The proposed combination of causal inference with a two-stage hurdle model jointly estimates the conditional average treatment effect on customer value and the purchase probability under treatment. The empirical results demonstrate that the consideration of treatment cost substantially increases campaign profit when used for customer targeting. The proposed causal hurdle model streamlines model building while achieving competitive campaign profit compared to the benchmark approaches.

Chapter 6 develops a framework for cost-efficient treatment randomization in randomized controlled trials. The fundamental problem of causal inference is that each individual can receive only a single treatment and that the hypothetical outcome under the treatment option that they did not receive remains unobservable. Causal inference must, therefore, rely on comparing the outcomes for groups of individuals with the same characteristics, where each group has received a different treatment. If the individuals in each group are not identical in the statistical sense, the treatment effect estimates will be biased. The confounding bias is a result of the difference in distribution between the treatment and control groups and can be caused, for example, by an existing targeting policy that assigns treatment based on customer attributes. Importantly, the confounding bias does not vanish with the collection of more data in the form of more observations or more covariates (Gordon et al., 2019).
Randomized controlled trials are an effective method to avoid confounding bias when estimating treatment effects. Randomization avoids confounding by replacing the existing targeting policy with random assignment of the marketing treatment. However, randomized treatment indiscriminately targets prospects who are deemed profitable or unprofitable based on their attributes. This makes experimental data collection using randomization costly. The unequal size of treatment and control groups in practice suggests that companies are aware of these costs (Diemert et al., 2018; Kane et al., 2014). This essay proposes to retain the existing prediction model during data collection and to introduce a stochastic component to the targeting policy instead of fully randomized treatment assignment. The stochastic policy is applied to the estimates of the prediction model and therefore poses no restriction on the model specification. Controlled randomization based on observed attributes allows full correction of the treatment and control group distributions using an established methodology designed for observational studies. Combining model-based customer targeting and randomized exploration reduces the cost of data collection and enables the continuous collection of data for model evaluation and updating. Continuous data collection is critical for non-disruptive experimentation and moni-

toring the performance of uplift models in deployment.

Chapter 7 bridges data collection from the perspective of the company and the perspective of the customer by quantifying the tradeoff between profit on personalized marketing and levels of customer privacy. The goal of personalized marketing is an additional profit on marketing spending that can only be achieved based on the collection of customer information, which has become a characterizing feature of the digital economy (Ansari & Mela, 2003). Existing research has focused on the monetary costs of data acquisition to the company. These costs include direct acquisition costs (e.g. Bolón-Canedo et al., 2014; Maldonado et al., 2017) and, increasingly, compliance during data collection and storage (Hand, 2018). In addition to legal compliance, the collection of personal data infringes on the customer interest of data privacy (McDermott, 2017).
This essay contributes to the literature by quantifying the cost of data privacy by analyzing the tradeoff between collecting more data to increase marketing effectiveness and collecting less or less critical data to preserve customer privacy. In the first step, we identify the levels of privacy risk attached to the information collected for customer targeting and propose privacy risk classes for the data available to online retailers. In the second step, we evaluate the effectiveness of customer targeting due to data with increasing privacy risk classification. The results suggest that session-based customer information is most informative for purchase prediction. Data aggregation over time requires persistent customer identification, but shows a substantial additional benefit to model performance. Within the boundaries of the study, information that could be used to identify individual customers shows little incremental value for purchase prediction. These results provide a nuanced challenge against the trend to indiscriminately collect customer information. The study provides a template for quantifying the cost of data security and privacy that serves as an example for practitioners to include an evaluation of customer privacy into managerial decisions on data collection.

The remaining chapters expand on customer privacy concerns in the particular context of personalized marketing in emails. A distinct trend of personalized marketing is the extent to which interaction with the customer can be initiated and monitored (Bonfrer & Drèze, 2009; Bujlow et al., 2017). Email tracking applies techniques from web tracking to a different communication channel to monitor reading behavior for customer relationship management (Hasouneh & Alqeed, 2010). Email tracking refers to the collection of data generated by the recipient's interaction with an email. This data is collected by embedding images into the email which are downloaded and rendered by the recipient's client when the email is opened. User identifiers inserted into the server request for the embedded image allow marketers to infer that an email was read by the customer, when it was read and how often. User identifiers inserted into referral links within the email further allow marketers to connect the user's email address and email reading behavior to the customer's interaction with the marketer's website.

Chapter 8 explores the privacy implications of email tracking and investigates the prevalence of email tracking in marketing communication. Email tracking has been discussed in the data privacy literature almost exclusively as an extension of web tracking to which it is technically

similar (Martin et al., 2003). For example, it is possible to infer information about the customer by analyzing the user agent string, which includes their device and operating system (Agosti & Di Nunzio, 2007). More personally, a customer's affiliation to a company or institution can be uncovered based on a reverse lookup of the IP address requesting the embedded image. Location-related information can be gathered using geolocation services (Poese et al., 2011). In contrast to web tracking, data collected through email tracking is not anonymous, since it is necessarily linked to the customer's email address that serves as a unique identifier. Personal identifiability and the potential for data sharing with third parties make email tracking a stronger privacy risk (Englehardt et al., 2018). Email tracking further exacerbates the privacy impact of existing web tracking by connecting the user's email address to cookie information stored on the device and leaking identifying information to third-party trackers (Englehardt et al., 2018). However, the extent to which email tracking data is utilized in email marketing is not well researched, although its use for monitoring is well-documented (Bonfrer & Drèze, 2009; Hasouneh & Alqeed, 2010).
To robustly identify email tracking in emails and investigate its prevalence in marketing communication, the study develops a methodology using controlled newsletter subscriptions. Embedded tracking images are identified by comparing the image embedding code received by identical subscribers. The comparison of 4,500 emails sent by the 100 largest companies in the United States, Britain and Germany shows that 51% contained at least one tracking image. Identifying and blocking these images can be achieved by several technical measures. The study identifies the classification of individual images within each email as the best measure to balance privacy protection with usability to the email recipient, based partially on the observation that 65% of the collected emails include company-specific rather than third-party tracking. This conclusion stands in contrast to the blacklist approaches favored in the detection of web tracking (Cormack, 2006).

Chapter 9 expands on Chapter 8 through experimental evaluation of the extent to which companies utilize the customer behavior information collected through email tracking. The study confirms the existence of tracking in a set of prominent newsletters and investigates the personalization of marketing communication related to differences in customer characteristics and reading behavior. To that end, twelve email accounts are created, each of which subscribes to a predefined set of newsletters from companies based in the United States, Britain and Germany. Each account simulates a different type of user with reading patterns that are systematically varied across accounts. The study finds that 13 out of 44 senders adjust their communication in response to user reading behavior, despite over 92% of the newsletter e-mails containing tracking images. Observed adjustments include sending newsletters at different times, increased or decreased mailing frequency and mobile-specific adjustments. The study further finds that only a single sender adapts the advertised products to the user under the caveat that no user behavior on the company website or related websites was simulated in the experiment. With regard to legal compliance, not all companies that adapt the mail-sending behavior state the purpose of their data collection in their privacy policy.

Chapters 10 and 11 develop a decision-support system to restore the data privacy of customers

by preventing email tracking. This system builds on a statistical classifier of images within the email to selective block tracking images.
Chapter 10 develops a model to identify the specific images in emails that are used for tracking. Identifying tracking images is a challenging task, because the available information is restricted to the code used to embed the image into the email and because the structure of this code is under the active control of the tracker. The first contribution of the study is the construction of a set of variables from HTML code that serves as input to the classifier. The input of the model is restricted to the HTML code used to embed the image in the email since loading the image content provides behavioral information to the tracker. These variables are devised to be computationally efficient and to generalize to structures of unseen tracking images. The second contribution of the study is the careful selection of variables that are resilient against changes in tracking structures and the development of a model for the robust classification of tracking images. A special concern is placed on the technical means of tracking providers to subvert detection efforts by actively manipulating the proposed variables. Third, using a selection of state-of-the-art classifiers, we test the predictive power of these features in a benchmarking experiment to clarify the effectiveness of model-based tracking detection. We evaluate the expected accuracy of the approach on test sets containing unseen emails after an increasing amount of time has passed and from previously unseen senders. This allows us to identify an optimal detection model and appraise the degree to which a model-based approach protects against email tracking in practice. The results show that robust model-based identification of tracking images is feasible with a minimum of inconvenience to the user.

Chapter 11 develops a solution to embed the detection system into a server-side architecture to scan and clean tracking images in incoming emails. Following the guidelines of Design Science (Hevner et al., 2004), its goal is the development and rigorous evaluation of an artifact based on the contributions of Chapter 8 to 10. The study conceptualizes, implements and evaluates a software extension to mail servers. This extension identifies tracking images in e-mails using the classifier developed in Chapter 11 and selectively replaces them with a placeholder image containing a warning message for the recipient. The anti-tracking server is developed as enterprise-grade software to generate knowledge on the design of server-based tracking solutions. It is flexibly extensible, highly scalable and ready to be applied in a production environment. The study provides extensive experimental evaluation in the dimensions of processing time, parallel requests and technical requirements for company-scale email servers. The results show that the proposed server-side solution can efficiently clean company-scale email traffic from tracking images. The solution is managerially relevant as it provides an off-the-shelf design for industry application and contributes to future research as it provides a modular online testbed for the evaluation of tracking detection algorithms.

Bibliography

Acquisti, A., Taylor, C., & Wagman, L. (2016). The economics of privacy. *Journal of Economic Literature*, *54*(2), 442–492. https://doi.org/10.1257/jel.54.2.442

Acquisti, A., & Varian, H. R. (2005). Conditioning prices on purchase history. *Marketing Science*, *24*(3), 367–381. https://doi.org/10.1287/mksc.1040.0103

Agosti, M., & Di Nunzio, G. M. (2007). Gathering and Mining Information from Web Log Files, In *Digital Libraries: Research and Development*, Berlin, Heidelberg, Springer. https://doi.org/10.1007/978-3-540-77088-6_10

Agrawal, R., Imielinski, T., & Swami, A. (1993). Database mining: A performance perspective. *IEEE Transactions on Knowledge & Data Engineering*, *5*(6), 914–925.

Anderson, R., & Moore, T. (2006). The economics of information security. *Science*, *314*(5799), 610–613. https://doi.org/10.1126/science.1130992

Ansari, A., & Mela, C. F. (2003). E-customization. *Journal of Marketing Research*, *40*(2), 131–145.

Ascarza, E. (2018). Retention futility: Targeting high risk customers might be ineffective. *Journal of Marketing Research*, *55*(1). https://doi.org/10.1509/jmr.16.0163

Ascarza, E., & Hardie, B. G. S. (2013). A joint model of usage and churn in contractual settings. *Marketing Science*, *32*(4), 570–590. https://doi.org/10.1287/mksc.2013.0786

Backiel, A., Baesens, B., & Claeskens, G. (2016). Predicting time-to-churn of prepaid mobile telephone customers using social network analysis. *Journal of the Operational Research Society*, *67*(9). https://doi.org/10.1057/jors.2016.8

Bahnsen, A. C., Aouada, D., & Ottersten, B. (2015a). Example-dependent cost-sensitive decision trees. *Expert Systems with Applications*, *42*(19), 6609–6619. https://doi.org/10.1016/j.eswa.2015.04.042

Bahnsen, A. C., Aouada, D., & Ottersten, B. (2015b). A novel cost-sensitive framework for customer churn predictive modeling. *Decision Analytics*, *2*(5), 1–15. https://doi.org/10.1186/s40165-015-0014-6

Bansal, G., Zahedi, F. M., & Gefen, D. (2015). The role of privacy assurance mechanisms in building trust and the moderating role of privacy concern. *European Journal of Information Systems*, *24*(6), 624–644. https://doi.org/10.1057/ejis.2014.41

Bauer, C. L. (1988). A direct mail customer purchase model. *Journal of Direct Marketing*, *2*(3), 16–24.

Blake, T., Nosko, C., & Tadelis, S. (2015). Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica*, *83*(1), 155–174. https://doi.org/10.3982/ECTA12423

Bodapati, A., & Gupta, S. (2004). A direct approach to predicting discretized response in target marketing. *Journal of Marketing Research*, *41*(1), 73–85.

Bolón-Canedo, V., Porto-Díaz, I., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2014). A framework for cost-based feature selection. *Pattern Recognition*, *47*(7), 2481–2489. https://doi.org/10.1016/j.patcog.2014.01.008

Bonfrer, A., & Drèze, X. (2009). Real-time evaluation of e-mail campaign performance. *Marketing Science*, *28*(2), 251–263. https://doi.org/10.1287/mksc.1080.0393

Bose, I., & Xi, C. (2009). Quantitative models for direct marketing: A review from systems perspective. *European Journal of Operational Research*, *195*(1), 1–16. https://doi.org/10.1016/j.ejor.2008.04.006

Bradlow, E. T., Gangwar, M., Kopalle, P., & Voleti, S. (2017). The role of big data and predictive analytics in retailing. *Journal of Retailing*, *93*(1), 79–95. https://doi.org/10.1016/j.jretai.2016.12.004

Bucklin, R. E., & Sismeiro, C. (2009). Click here for internet insight: Advances in clickstream data analysis in marketing. *Journal of Interactive Marketing*, *23*(1), 35–48. https://doi.org/10.1016/j.intmar.2008.10.004

Bujlow, T., Carela-Espanol, V., Sole-Pareta, J., & Barlet-Ros, P. (2017). A survey on web tracking: Mechanisms, implications, and defenses. *Proceedings of the IEEE*, *105*(8), 1476–1510. https://doi.org/10.1109/JPROC.2016.2637878

Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, A. (2004). Ensemble Selection from Libraries of Models, In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, New York, ACM.

Chan, T. Y., Wu, C., & Xie, Y. (2011). Measuring the lifetime value of customers acquired from Google search advertising. *Marketing Science*, *30*(5), 837–850.

Chen, Z.-Y., Fan, Z.-P., & Sun, M. (2015). Behavior-aware user response modeling in social media: Learning from diverse heterogeneous data. *European Journal of Operational Research*, *241*(2), 422–434. https://doi.org/10.1016/j.ejor.2014.09.008

Cormack, G. V. (2006). Email spam filtering: A systematic review. *Foundations and Trends® in Information Retrieval*, *1*(4), 335–455.

De Caigny, A., Coussement, K., De Bock, K. W., & Lessmann, S. (2019). Incorporating Textual Information in Customer Churn Prediction Models Based on a Convolutional Neural Network. *International Journal of Forecasting*, *In Press*.

Devriendt, F., Moldovan, D., & Verbeke, W. (2018). A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. *Big Data*, *6*(1), 13–41. https://doi.org/10.1089/big.2017.0104

Diemert, E., Betlei, A., Renaudin, C., & Amini, M.-R. (2018). A Large Scale Benchmark for Uplift Modeling, In *Proceedings of the AdKDD and TargetAd Workshop, KDD*, London, United Kingdom, ACM.

Dubé, J.-P., Fang, Z., Fong, N., & Luo, X. (2017). Competitive price targeting with smartphone coupons. *Marketing Science*, *36*(6), 944–975. https://doi.org/10.1287/mksc.2017.1042

Elkan, C. (2001). The Foundations of Cost-Sensitive Learning (B. Nebel, Ed.). In B. Nebel (Ed.), *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann.

Englehardt, S., Han, J., & Narayanan, A. (2018). I Never Signed Up For This! Privacy Implications of Email Tracking, In *Proceedings on Privacy Enhancing Technologies*.

Finlay, S. (2010). Credit scoring for profitability objectives. *European Journal of Operational Research*, *202*(2), 528–537.

Fuller, R. M., & Dennis, A. R. (2009). Does fit matter? The impact of task-technology fit and appropriation on team performance in repeated tasks. *Information Systems Research*, *20*(1), 2–17. https://doi.org/10.1287/isre.1070.0167

Gabel, S. (2019). *One-to-One Marketing in Grocery Retailing* (Dissertation). Humboldt-Universität zu Berlin. Berlin.

Glady, N., Baesens, B., & Croux, C. (2009). Modeling churn using customer lifetime value. *European Journal of Operational Research*, *197*(1), 402–411.

Goldfarb, A., & Tucker, C. (2011). Online display advertising: Targeting and obtrusiveness. *Marketing Science*, *30*(3), 389–404. https://doi.org/10.1287/mksc.1100.0583

Goldfarb, A., & Tucker, C. E. (2011). Privacy Regulation and Online Advertising. *Management Science*, *57*(1), 57–71. https://doi.org/10.1287/mnsc.1100.1246

Gordon, B. R., Zettelmeyer, F., Bhargava, N., & Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Science*, *38*(2), 193–364. https://doi.org/10.1287/mksc.2018.1135

Grbovic, M., Radosavljevic, V., Djuric, N., Bhamidipati, N., Savla, J., Bhagwan, V., & Sharp, D. (2015). E-commerce in Your Inbox: Product Recommendations at Scale, In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2015)*, Sydney, Australia. https://doi.org/10.1145/2783258.2788627

Gubela, R. M., Bequé, A., Gebert, F., & Lessmann, S. (2019). Conversion uplift in e-commerce: A systematic benchmark of modeling strategies. *International Journal of Information Technology & Decision Making*, *18*(3), 747–791.

Hand, D. J. (2018). Aspects of data ethics in a changing world: Where are we now? *Big Data*, *6*(3), 176–190. https://doi.org/10.1089/big.2018.0083

Hansotia, B., & Rukstales, B. (2002). Incremental value modeling. *Journal of Interactive Marketing*, *16*(3), 35–46. https://doi.org/10.1002/dir.10035

Hartemo, M. (2016). Email marketing in the era of the empowered consumer. *Journal of Research in Interactive Marketing*, *10*(3), 212–230. https://doi.org/10.1108/JRIM-06-2015-0040

Hasouneh, A. B. I., & Alqeed, M. A. (2010). Measuring the effectiveness of e-mail direct marketing in building customer relationship. *International Journal of Marketing Studies*, *2*(1), 48–64. https://doi.org/10.5539/ijms.v2n1p48

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105.

Hitsch, G. J., & Misra, S. (2018). Heterogeneous Treatment Effects and Optimal Targeting Policy Evaluation. *SSRN*.

Hormozi, A. M., & Giles, S. (2004). Data mining: A competitive weapon for banking and retail industries. *Information Systems Management*, *21*(2), 62–71.

Ishigaki, T., Terui, N., Sato, T., & Allenby, G. M. (2018). Personalized market response analysis for a wide variety of products from sparse transaction data. *International Journal of Data Science and Analytics*. https://doi.org/10.1007/s41060-018-0099-9

Jacobs, B. J. D., Donkers, B., & Fok, D. (2016). Model-based purchase predictions for large assortments. *Marketing Science*, *35*(3), 389–404. https://doi.org/10.1287/mksc.2016.0985

Kane, K., Lo, V. S. Y., & Zheng, J. (2014). Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. *Journal of Marketing Analytics*, *2*(4), 218–238. https://doi.org/10.1057/jma.2014.18

Kim, Y. S., Lee, H., & Johnson, J. D. (2013). Churn management optimization with controllable marketing variables and associated management costs. *Expert Systems with Applications*, *40*(6), 2198–2207. https://doi.org/10.1016/j.eswa.2012.10.043

Kim, Y. S., & Moon, S. (2012). Measuring the success of retention management models built on churn probability, retention probability, and expected yearly revenues. *Expert Systems with Applications*, *39*(14), 11718–11727. https://doi.org/10.1016/j.eswa.2012.04.048

Kim, Y. S., Street, W. N., Russell, G. J., & Menczer, F. (2005). Customer targeting: A neural network approach guided by genetic algorithms. *Management Science*, *51*(2), 264–276.

Kohavi, R., Mason, L., Parekh, R., & Zheng, Z. (2004). Lessons and challenges from mining retail e-commerce data. *Machine Learning*, *57*(1), 83–113.

Kumar, V., Venkatesan, R., Bohling, T., & Beckmann, D. (2008). The power of CLV: Managing customer lifetime value at IBM. *Marketing Science*, *27*(4), 585–599.

Lemmens, A., & Gupta, S. (2017). Managing Churn to Maximize Profits. *Social Science Research Network*, 2964906. https://doi.org/10.2139/ssrn.2964906

Lian, S., Cha, T., & Xu, Y. (2019). Enhancing geotargeting with temporal targeting, behavioral targeting and promotion for comprehensive contextual targeting. *Decision Support Systems*, *117*, 28–37. https://doi.org/10.1016/j.dss.2018.12.004

Lo, V. S. Y. (2002). The true lift model: A novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter*, *4*(2), 78–86.

Maldonado, S., Bravo, C., López, J., & Pérez, J. (2017). Integrated framework for profit-based feature selection and SVM classification in credit scoring. *Decision Support Systems*, *104*, 113–121. https://doi.org/10.1016/j.dss.2017.10.007

Martens, D., Provost, F., Clark, J., & Junqué de Fortuny, E. (2016). Mining massive fine-grained behavior data to improve predictive analytics. *MIS Quarterly*, *40*(4), 869–888. https://doi.org/10.25300/misq/2016/40.4.04

Martin, D., Wu, H., & Alsaid, A. (2003). Hidden surveillance by web sites: Web bugs in contemporary use. *Communications of the ACM*, *46*(12), 258. https://doi.org/10.1145/953460.953509

Mayer, J. R., & Mitchell, J. C. (2012). Third-Party Web Tracking: Policy and Technology, In *IEEE Symposium on Security and Privacy*, San Francisco, CA, USA, IEEE. https://doi.org/10.1109/SP.2012.47

McDermott, Y. (2017). Conceptualising the right to data protection in an era of Big Data. *Big Data & Society*, *4*(1), 1–7. https://doi.org/10.1177/2053951716686994

Moeyersoms, J., & Martens, D. (2015). Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector. *Decision Support Systems*, *72*, 72–81. https://doi.org/10.1016/j.dss.2015.02.007

Nofer, M., Hinz, O., Muntermann, J., & Roßnagel, H. (2014). The economic impact of privacy violations and security breaches: A laboratory experiment. *Business & Information Systems Engineering*, *6*(6), 339–348. https://doi.org/10.1007/s12599-014-0351-3

Park, C. H., & Park, Y.-H. (2016). Investigating purchase conversion by uncovering online visit patterns. *Marketing Science*, *35*(6), 894–914. https://doi.org/10.1287/mksc.2016.0990

Poese, I., Uhlig, S., Kaafar, M. A., Donnet, B., & Gueye, B. (2011). IP geolocation databases: Unreliable? *ACM SIGCOMM Computer Communication Review*, *41*(2), 53. https://doi.org/10.1145/1971162.1971171

Pollach, I. (2011). Online privacy as a corporate social responsibility: An empirical study. *Business Ethics: A European Review*, *20*(1), 88–102. https://doi.org/10.1111/j.1467-8608.2010.01611.x

Rossi, P. E., Allenby, G. M., & McCulloch, R. E. (2005). *Bayesian statistics and marketing*. Hoboken, NJ, Wiley.

Ruiz, F. J. R., Athey, S., & Blei, D. M. (2017). SHOPPER: A Probabilistic Model of Consumer Choice with Substitutes and Complements. *CoRR*, *abs/1711.03560*.

Rzepakowski, P., & Jaroszewicz, S. (2012). Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, *32*(2), 303–327. https://doi.org/10.1007/s10115-011-0434-0

Sahni, N. S., Wheeler, S. C., & Chintagunta, P. (2018). Personalization in email marketing: The role of noninformative advertising content. *Marketing Science*, *37*(2), 236–258. https://doi.org/10.1287/mksc.2017.1066

Sahni, N. S., Zou, D., & Chintagunta, P. K. (2016). Do targeted discount offers serve as advertising? Evidence from 70 field experiments. *Management Science*, *63*(8), 2688–2705. https://doi.org/10.1287/mnsc.2016.2450

Shaffer, G., & Zhang, Z. J. (2002). Competitive one-to-one promotions. *Management Science*, *48*(9), 1143–1160.

Shaw, M. J., Subramaniam, C., Tan, G. W., & Welge, M. E. (2001). Knowledge management and data mining for marketing. *Decision Support Systems*, *31*(1), 127–137.

Stange, M., & Funk, B. (2014). Real-time advertising. *Business & Information Systems Engineering*, *6*(5), 305–308. https://doi.org/10.1007/s12599-014-0346-0

Stripling, E., vanden Broucke, S., Antonio, K., Baesens, B., & Snoeck, M. (2018). Profit maximizing logistic model for customer churn prediction using genetic algorithms. *Swarm and Evolutionary Computation*, *40*, 116–130.

Tian, L., Alizadeh, A. A., Gentles, A. J., & Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, *109*(508), 1517–1532. https://doi.org/10.1080/01621459.2014.951443

Van den Poel, D., & Prinzie, A. (2005). Constrained optimization of data-mining problems to improve model performance: A direct-marketing application. *Expert Systems with Applications*, *29*(3), 630–640. https://doi.org/10.1016/j.eswa.2005.04.017

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, *218*(1), 211–229. https://doi.org/10.1016/j.ejor.2011.09.031

Yang, S., Lu, S., & Lu, X. (2014). Modeling competition and its impact on paid-search advertising. *Marketing Science*, *33*(1), 134–153. https://doi.org/10.1287/mksc.2013.0812

Zhang, Y., Bradlow, E. T., & Small, D. S. (2015). Predicting customer value using clumpiness: From RFM to RFMC. *Marketing Science*, *34*(2), 195–208. https://doi.org/10.1287/mksc.2014.0873

Zhu, B., Baesens, B., Backiel, A., & vanden Broucke, S. K. L. M. (2017). Benchmarking sampling techniques for imbalance learning in churn prediction. *Journal of the Operational Research Society*. https://doi.org/10.1057/s41274-016-0176-1

Chapter 2

Changing Perspectives: Using Graph Metrics to Predict Purchase Probabilities

Publication
Baumann, A., Haupt, J., Gebert, F., & Lessmann, S. (2018). Changing perspectives: Using graph metrics to predict purchase probabilities. Expert Systems with Applications, 94, 137–148.

Abstract
The prediction of user online behavior (next clicks, repeat visits, purchases, etc.) is a well-studied subject in research. Prediction models typically rely on clickstream data that is captured during the visit of a website and embodies user agent-, path-, time- and basket-related information. The aim of the paper is to propose an alternative approach to extract auxiliary information from the website navigation graph of individual users and to test the predictive power of this information. Using two real-world large datasets of online retailers we develop an approach to construct within session graphs from clickstream data and demonstrate the relevance of corresponding graph metrics to predict purchases.

2.1 Introduction

The e-commerce sector is responsible for a substantial fraction of firm revenues. Annual turnover was 1,336 billion US dollars in 2014 and is predicted to have reached 2,050 billion US dollars in 2016 (eMarketer, 2019b). However, given that growth rates are expected to decline in the future (eMarketer, 2019a), e-commerce shops need to find ways to defend market shares in an increasingly competitive environment. One strategy to do so is to increase purchasing amounts and/or frequencies of existing customers. Important determinants of (re-)purchase intention in online shopping are trust, service quality (Hong & Kim, 2012) and the satisfaction a user experiences during the online shopping process (H. Lee et al., 2009). To offer a richer user experience and increase visitors' (re-)purchase intentions, understanding customer online behavior is crucial (e.g. Pai et al., 2014). To gain such insight and to anticipate user actions, the analysis of clickstream data has been widely adopted in the literature (C. H. Park & Park, 2016; Van den Poel & Buckinx, 2005).

However, previous work in the field has not examined the potential of graph theory to gather auxiliary information from clickstream data and increase the accuracy of behavior prediction

models. Graphs are a methodological approach originating from network theory. They consist of nodes and edges, which connect nodes. Graph-based approaches have been used in various fields and have been proven to be helpful for various tasks, for example to predict connections in the social networking context (He et al., 2015), to detect money laundering activities (Colladon & Remondi, 2017), for personalized recommendations (Shams & Haratizadeh, 2017) and for customer churn prevention (Óskarsdóttir et al., 2017). Given the success of graph-based predictors in these and other applications, the objective of our paper is to test their potential for online behavior prediction based upon clickstream data.

We contribute to literature as follows: First, we propose an approach to derive graphs out of user sessions based on clickstream data. Second, we calculate graph metrics and examine their pairwise dependency in terms of correlation. Third, we assess how they perform as a means to predict customer behavior in online contexts.

The remainder of the paper is structured as follows. First, we give an overview on relevant literature to clarify the research gap the paper strives to. Afterwards, we present our methodology and how we derive clickstream graphs in particular. We then summarize the resulting data, before presenting empirical results. Last, we summarize our findings.

2.2 Related Work

Much literature considers the use of clickstream data for customer online behavior prediction. Prediction targets range from conversions in purchase prediction (Van den Poel & Buckinx, 2005), whether visitors redeem incentives (Pai et al., 2014) or complete specific tasks such as putting an item into a basket (Kalczynski et al., 2006; Sismeiro & Bucklin, 2004), over navigational behavior prediction (Montgomery et al., 2004, e.g., the next web site) to classifying visitors into interest groups such as whether a user's site visiting intention is informational or transactional (Moe, 2003).

Table 2.1 summarizes related work, which we categorize according to the target of prediction into navigational behavior (NB), user classification (UC) and conversion (PC) prediction, where PC is the prevailing target in prior work (i.e., 23 out of 34 studies fall in this category).

Table 2.1 also shows the types of features (i.e., covariates) which the studies employ for predictive modelling. In particular, we categorize the features into six groups. All categories except demographics are based on clickstream data. The first three groups – time, page and monetary – draw inspiration from the well-known concept of recency, frequency and monetary value analysis (Zhang et al., 2015). Recency and frequency consist of aspects such as time on page and last website visit (*Time*), whereas monetary comprises historical purchase behavior derived from preceding clickstream sessions and current basket information (*Monetary*). Frequency refers to the path traversal and categories of pages visited, counting how often each has been visited (*Page*). In addition, we consider behavior related variables (*Page Interaction*), such as basket interaction, click on page and scroll on page events to capture user-centered feature categories, which resolve around behavioral aspects besides the website path that a user traverses. The

Table 2.1: Overview of feature categories used in research (NB: Navigational Behaviour, UC: User Classification, PC: Conversion)

Reference	Dependent Variable	Feature Category					
		Page	Time	Monetary	Page Interaction	Demo-graphics	Graph/ Similarity
Anitha 2010	NB	x					
Antonellis et al. 2009	UC	x					
Banerjee & Ghosh 2001	UC	x	x				x
Berka & Labsky 2007	NB	x					
Byeon 2013	PC	x	x				x
Chan et al. 2014	PC	x			x	x	
Girija & Kavitha 2013	NB	x					
Iwanaga et al. 2016	PC	x	x				
Jiang et al. 2012	PC	x	x				
Kalczynski et al. 2006	PC	x					x
Lee et al. 2010	PC	x	x				
Lu et al. 2005	UC	x					
Moe 2003	UC	x	x				
Moe & Fader 2004	PC		x	x			
Moe et al. 2002	PC	x	x				
Montgomery et al. 2004	NB	x					
Gündüz & Özsu 2003	NB	x	x				x
Padmanabhan et al. 2006	PC / Re-visit	x	x	x		x	
Pai et al. 2014	UC	x	x				
Panagiotelis et al. 2014	PC	x	x	x			
Park et al. 2008	UC	x					
Park & Park 2016	PC	x					
Pitman & Zanker 2010	PC	x			x		
Sarwar et al. 2015	PC	x	x	x			
Sato & Asahi 2012	PC (day)	x		x			
Senecal et al. 2014	UC	x	x		x		
Sismeiro & Bucklin 2004	PC	x	x		x		
Stange & Funk 2015	PC	x	x	x	x		
Suh et al. 2004	PC	x	x				
Van den Poel & Buckinx 2005	PC	x	x	x		x	
Vroomen et al. 2005	PC		x	x	x	x	
Wu et al. 2005	PC	x					
Zhao et al. 2016	PC	x	x	x			
Zheng et al. 2003	PC	x	x	x		x	

feature category *Demographics* consists of variables that capture user characteristics, which are not related to the website itself and thus not part of clickstream data, such as gender and geographic-related information. The last category captures studies which use graphs as a tool to derive features for their models used (*Graph / Similarity*).

Only four of the 34 studies, which base their analysis on clickstream, use a graph-based approach. In view of Table 2.1, it becomes evident that combining predictive modelling with graph-based features has been rare so far. Byeon (2013) generates for each user and each session a bi-partite graph (i.e. a graph with two different types of nodes), where the nodes represent a specific webpage, which a user visits during her session, and the category to which the webpage belongs, respectively. Each graph facilitates the calculation of summary statistics (i.e., density), which Byeon (2013) employs to predict whether a user session leads to a pur-

chase using a logistic regression classifier. In comparison to classical clickstream features (e.g. total number of clicks, total visit time), the graph density feature provides encouraging results, suggesting that it is a good predictor of purchase intention.

Kalczynski et al. (2006) predict whether a specific website task has been completed successfully. To achieve this, they focus on navigational complexity. The authors construct an experimental setup where users are asked to browse a website to conduct an artificial purchase. The website data is based on five different datasets which they use to derive graphs out of user journeys on a website. The authors then calculate a set of graph measures, some of which are based on specific website characteristics. Finally, they employ logistic regression to predict online task completion and conversion in particular.

Other approaches aiming at user classification do so via clustering using graph-based approaches to be able to build similarity graphs, connecting users which behave similar on websites. Banerjee and Ghosh (1997) use clickstream data to create a similarity graph, which connects users who display similar website usage behavior. First, they select pair-wise user sessions and compare them in terms of path and time dimensions to derive a similarity score. They then construct a weighted graph with nodes representing users which are pairwise connected once the weight reaches a specific threshold. The weight represents the similarity between two users. The similarity graph serves as input to a graph-based clustering method to derive user groups.

A similar approach is applied by Gündüz and Özsu (2003), who construct a similarity graph to apply a graph-based clustering method. Their graph is based on path and time aspects associated with a user journey on a website. The aim of graph construction and clustering is to predict the next website request.

The review of related work suggests that a comprehensive study, which systematically assesses the predictive value of a broad set of graph metrics is lacking. Building on the work of Byeon (2013) and Kalyznski et al. (2015) to predict purchase intention/conversion, we contribute toward closing this research gap in that we i) develop a way to derive a graph from clickstream data, ii) consider a much richer set of graph metrics, iii) employ real-life data, and iv) use a state-of-the-art prediction algorithms (random forest, gradient boosting machine) alongside logistic regression.

2.3 Methodology

The following sections explain our approaches to create clickstream graphs and derive corresponding graph metrics as input for predictive modeling.

2.3.1 Clickstream and Graph Construction

Clickstream is defined as the path which a website user traverses when visiting a number of websites (Bucklin et al., 2002) and consists of sessions each of which represent a single visit of a user on a website. Each session consists of an arbitrary number of page views, which are the webpages the user visits during a session. Specific behavior can be performed on a webpage such

as click, scroll and basket events. Furthermore, single webpages are visited for a specific amount of time. So far, the representation of clickstream in the form of a graph has been established mainly for visualization purposes (e.g. Kitts et al., 2002). Using clickstream data, we construct a graph for each session of a user to be able to derive covariates for purchase prediction. In general, a graph $G = (V, E)$ consists of a set of nodes V which are pair-wise connected via edges E. The edges are either directed or undirected. Each graph can be represented as a $n \times n$ adjacency matrix whose elements a_{ij} are set to one if node n_i and n_j are connected, and zero otherwise.

The user session graphs applied in this paper are constructed in the following way: Each node represents a specific website a user has visited during the session. For each page view, we create a new node if it does not already exist in the session graph (i.e., the user has not visited the page before). We connect two nodes with an edge (i.e., between two pages), if the user visits them successively. The edges are directed to capture the specific order in which webpages are visited. Due to incremental node insertion, the session graph grows successively during the users' journey on the website. This technique is known as "clipping at every click" (VanderMeer et al., 2000), meaning that we calculate for every page view a new graph and its underlying graph metrics to capture the user sessions' characteristics in an incremental manner. Figure 2.1 shows an example of this approach where a session of a user is represented as a graph structure which is updated at every page view, i.e., every webpage the user visits during her journey on the website. Furthermore, as an example the incremental calculation of a graph metric (i.e., average in-degree, which is the average number of edges converging to a node) is shown, which is re-calculated at each page view of a user.

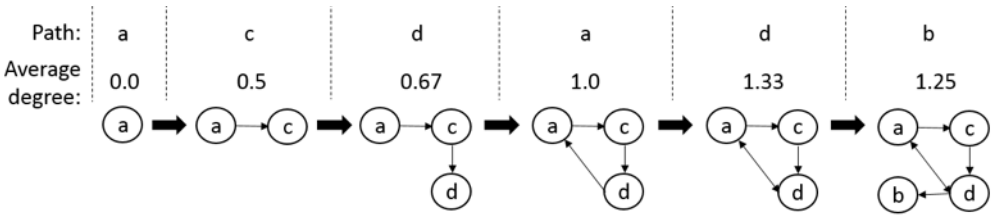


Figure 2.1: Example of a graph inference of a user session based on clickstream data

Based upon Kalczynski et al. (2006), we assume that the structure of a graph represents user behavior, which motivates examining the potential of graph metrics to predict conversion. More specifically, a graph based on clickstream data grounds on the explicit user click behavior and captures the path traversal of a user on a website. This behavior is a result of the user's goal in visiting the website and changes observably with user intention. To illustrate this, exemplary session graphs for three types of user behavior are shown in Figure 2.2.

The left graph shows a direct clickstream path, traversing from one page to another without returning behavior. This indicates a high degree of goal-orientation, which can be associated with both informational (the sought-after piece of information was found) or transactional (the desired product was bought) behavior. The middle graph illustrates a customer looking at various products of two types of product categories before deciding which is of further interest. This is a typical comparison behavior. The right-most graph depicts broad browsing behavior.

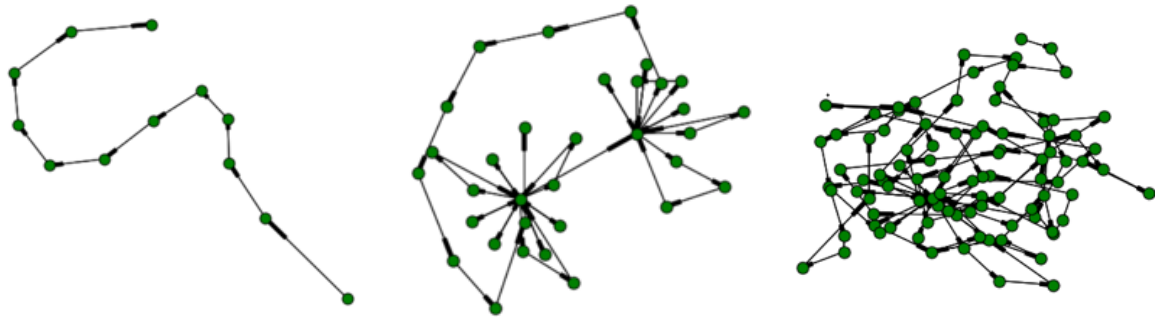


Figure 2.2: Graph visualizations of user sessions representing different types of user behavior

It contains several central nodes signifying, for example, the overview pages of search results to which the user returns after looking at a sequence of products. These and other graph structures are captured by the graph metrics we apply and reduce to input suitable for predictive modelling.

2.3.2 Selected Graph Metrics

Several metrics have been established in graph theory to represent the characteristics of a graph. These metrics focus on a specific node, the n-hop neighborhood or the whole graph. Given our objective to clarify the relative predictiveness of different graph metrics in combination with the scarcity of prior work on graph-based online behavior prediction, it is not clear which graph metrics are the most informative. Therefore, we use the Python framework NetworkX (Schult, 2008) to create a large number of alternative metrics and compare their relative predictiveness empirically. To that end, we use a set of 23 graph measures in total (see Table 2.2). Metrics focusing on characteristics of single nodes are computed for all nodes in the graph and then averaged. We focus on structural, centrality- and distance-based metrics since they are able to describe the relative importance of graph elements in the network and the general structure of a graph. This in turn might be indicative of the users' intention behind the website visit.

Structural measures describe the general construction of a graph. The most basic concepts are the total number of nodes *N,* capturing the total number of unique webpages which a user visits, and the total number of edges *M,* being the number of directed and unique path traversals from one page to another. The number of circles and self-loops in a graph accounts for switching behavior of a user returning to a previously visited page or the same page, respectively. Related to these measures is flow hierarchy which states the proportion of edges not being part of a cycle. Transitivity indicates whether a user resolves around a specific subset of webpages such as switching between different products to choose from. Unlike transitivity, which focuses on the neighborhood of a node, density can be used as an additional measure of purposefulness of a session, since small values indicate a step-by-step list of pages without circularity or jump-backs hinting at more goal-oriented user behavior. Last, high values of the metric average node connectivity signal an interwoven connection between a considerable number of nodes and therefore a non-structured browsing behavior of users.

Distance measures relate to the broadness of a graph structure. The average shortest path length and the related metrics eccentricity, diameter, radius, center and periphery give an

indication how diverse the user traverses through the website. Their intuitive interpretation is that for high values of, e.g., the average shortest path length, the user rather looks at unique webpages one after another without the occurrence of returning behavior. Small values signal returning behavior to formerly visited pages such as overview and search result pages.

Centrality measures describe the importance of nodes in terms of how central they are located in the graph structure. For example, a particular node can be described by its degree centrality which can be interpreted as the amount to which users return to a specific page from varying other pages. Betweenness centrality for both nodes and edges measures whether there are bridging elements in the network structure, such as specific overview pages which user frequently returns to in order to access other webpages. Therefore, both concepts can be seen as examples for comparison behavior of a user. Eigenvector, katz and pagerank centrality indicate whether there is a wide choice of disjoint paths, where high average values indicate an interwoven structure of several important nodes. The intuition of closeness centrality and closeness vitality is that both measures assume high values if a node is located central in the whole network. For example, this applies to specific webpages a user has visited several times during the whole session. This extends the notion that there may be a specific number of pages that are central in the clickstream. The summary statistics for the graph metrics of both shops are provided in the appendix.

2.3.3 Prediction Model Training and Assessment

We use prediction models to forecast whether a user session leads to a purchase. We set this target variable to one for all page views in a session if the user conducts a purchase during the session; and to zero otherwise. All predictive variables, i.e. the selected graph metrics and control variables introduced below, are normalized to their standard score to facilitate the interpretation of coefficients for the linear model in terms of their standard deviation from the mean.

We perform out-of-time validation and split our datasets sequentially into training and set; according to the month of the session. Data from September is used as training set whereas data from August is used as test set, resulting in an approximate split of 6:4 between training and test data. Out-of-sample in combination with out-of-time validation is commonly used in benchmarking studies to understand model performance in marketing (Berry & Linoff, 2011, p.72) or credit scoring (Sobehart et al., 2000), where models are required to be stable over time. This is especially relevant in the e-commerce setting, since we want to test whether our model is able to predict the focal behavior for a different time period than the one in which the model was trained. The out-of-time validation approach is thus stricter in analyzing the performance of the model compared to randomized out-of-sample testing within the same period. We tune the meta-parameters of the prediction models introduced below by means of 5-fold cross validation on the training set. Since our data is highly imbalanced, we additionally applied synthetic minority over-sampling (SMOTE) (Chawla et al., 2002), which creates artificial data points based upon the characteristics of a real observation of the minority class and its direct neighborhood to create a balanced dataset.

Table 2.2: Overview of the applied graph metrics (asterik marks metrics removed from the final feature set due to multicollinearity)

Category	Metric	Feature	Description
Structure (8 metrics)	Number of nodes	NumberNodes*	Total number of nodes in the graph.
	Number of edges	NumberEdges*	Total number of edges in the graph.
	Number of cycle	NumberCircles	Total number of circles in the graph.
	Number of self-loops	SelfLoops	Total number of self-loops in the graph.
	Flow hierarchy	FlowHierarchy	Proportion of edges not being part of a cycle.
	Transitivity	Transitivity	The number of triangles in the graph divided by the maximum possible number of triangles.
	Density	Density	The sparseness in terms of connectivity for the whole graph.
	Mean node connectivity	NodeConnectivity*	Average number of nodes for each distinct node pair that must be removed from the network in order to disconnect them.
Distance (6 metrics)	Mean shortest path length	ShortestPath*	The average of the shortest path length for all distinct node pairs in the graph.
	Mean eccentricity	Eccentricity*	Mean of the longest shortest path for each single node in the graph.
	Diameter	Diameter*	The maximum eccentricity for the whole graph.
	Radius	Radius	The minimum eccentricity for the whole graph.
	Center	Center	Number of nodes with an eccentricity value equal to the radius.
	Periphery	Periphery	Number of nodes with an eccentricity value equal to the diameter.
Centrality (9 metrics)	Mean in-degree/ average out-degree	Degree*	Mean of the number of edges converging from/to a node.
	Mean neighbor degree	NeighborDegree	The average of the neighbor degree for each distinct node in the graph.
	Mean closeness centrality	Closeness*	The average closeness, i.e. centrality of all nodes in the graph.
	Mean closeness vitality	Vitality	The average change in closeness for all nodes if successively one node is removed from the graph.
	Mean node betweenness centrality	NodeBetweenness	Importance of a node in terms of number of shortest paths passing through this node.
	Mean edge betweenness centrality	EdgeBetweenness	Importance of an edge in terms of number of shortest paths passing through this edge.
	Mean eigenvector centrality	Eigenvector	Different measures to compute the centrality of a node based on the adjacency matrix of the graph considering the linkage structure of the direct neighborhood of a node and partially a node's own edge structure.
	Mean katz centrality	Katz*	
	Mean pagerank centrality	Pagerank*	

We select three different classification algorithms, a generalized linear logistic regression model (GLM) and two nonlinear tree-based models, which are random forest (RF) and gradient boosting tree ensemble (GBT). We motivate the choice of logistic regression by its use in previous work (Byeon, 2013; Kalczynski et al., 2006). RF is chosen due its high performance in several forecasting benchmarks (e.g. Lessmann et al., 2015). We apply GBT as a third method, because recent studies have found them to perform superior in similar classification tasks when compared to GLM and RF (Fitzpatrick & Mues, 2016). All models have the advantage that they are interpretable to a degree, which we use to examine the relative predictiveness of alternative graph metrics. The coefficients of logistic regression are interpretable in direction and size and allow significance testing. The RF and GBT classifier provides variable importance scores, which also indicate the predictiveness of a variable (Breiman, 2001).

We provide two measures of prediction performance. We evaluate the models build on the respective variable sets using the area-under-the-precision-recall-curves (AUC-PR). The AUC-PR is commonly applied as a single-value metric similar to the area-under-the-ROC-curve (AUC) (Fawcett, 2006) for model evaluation in case of imbalanced datasets (Takaya Saito & Marc Rehmsmeier, 2015). The PR curve is constructed through pairwise plotting of precision and recall pairs at different classification thresholds, where recall is the proportion of observations predicted to be positive (i.e., purchase) in relation to all positive observations and precision is the rate of predictions that are correct. In general, the higher the value of AUC-PR, the better the model discriminates between the two classes. The second measure we apply is the lift index, which is a popular performance indicator for targeting models (Ling & Li, 1998). Under some assumption, lift is directly connected to the profitability of a targeting model (Martens et al., 2016; Piatetsky-Shapiro & Masand, 1999), which further motivates the choice of this performance indicator. The lift is based on a list of customer ordered according to their model-estimate conversion probability. In our case, the lift is defined as the share of hits, i.e. purchasers, in the top segment of $0<\theta<1$ of customers sorted by predicted purchase probability divided by the expected number of buyers in a random sample. More formally, lift L_{d} is defined as:

$$L_d = \frac{\hat{\pi}_d}{\hat{\pi}} \tag{2.1}$$

with $\hat{\pi}_d$ denoting the fraction of purchasers among the top-d customers and $\hat{\pi}$ the prior probability of purchase, the lift assesses the degree to which a prediction model improves over a random benchmark.

To be able to assess the performance of our graph-based methodology in comparison to standard approaches, we will use an additional second feature set originating from the standard approach of feature extraction from clickstream (Table 2.3). Related to Table 2.1, we will use covariates of different categories such as *Page* and *Time*.

Table 2.3: Overview of traditional features in comparison to our graph approach

Feature	Description
SessionOverview	Number of pages of type 'overview'/'product'/'sale'/'search' in session.
SessionProduct	
SessionSale	
SessionSearch	
TabVisible	Is the tab currently visible?
Weekday	The weekday the session was started (1 – 7).
DayOfMonth	Day of the month (1 – 31) the session starts.
SessionStartHour	Hour of the session start (morning - midday - evening - night).
TimeOnPage	Time spent on page.
SessionTime	Total time of session.
PageVisitedBefore	Indicator whether the page has been visited before in the session.
Browser	The type of the browser the client uses.
ScreenSize	The screen size resolution of the visitor.
WindowSize	The window resolution of the visitor.
LocationZip	The zip code area of the city the user accesses the website from.
MajorCity	Indicator whether the website access happens from a major city.

2.4 Empirical Results

Based on the methodology discussed above, we report our empirical results in three steps. First, we will take a detailed look at the correlation among the graph measures applied. Second, we analyze the performance of the tested classifiers based on AUC-PR and the lift measure. Last, we will investigate the different graph measures in order to better understand their impact on the predictive accuracy.

2.4.1 Dataset Description

We use a two-month period of clickstream data of two large online retailers selling clothing and footwear, respectively. The data was collected from August to September 2015 and contains information such as identifiers (e.g., user id and session id), geographic- and user-based information (e.g., user agent) as well as path-, time- and behavioral-related information with regard to a customer's journey on the respective website.

In the first step, we clean the data by deleting incomplete sessions and dismissing user sessions with less than four page views. Those sessions are referred to as bouncers which have no interest in the website in itself or generally to conduct a purchase. Furthermore, at least four clicks are necessary to complete the purchase process. With regard to potential bot elimination from the dataset we exclude one outlying user sessions with a length of 550 views, which we assume to be the product of automated website access.

The descriptive statistics of the final datasets are shown in Table 2.4. In total, the first shop contains 58,545 unique users performing a total of 692,975 page views. Of all 80,184 sessions, 4,256 sessions (approx. 5.31%) result in a purchase by a user. The second shop has a lower visitor count of 18,759 users who account for 32,850 sessions and 475,500 page views. Looking

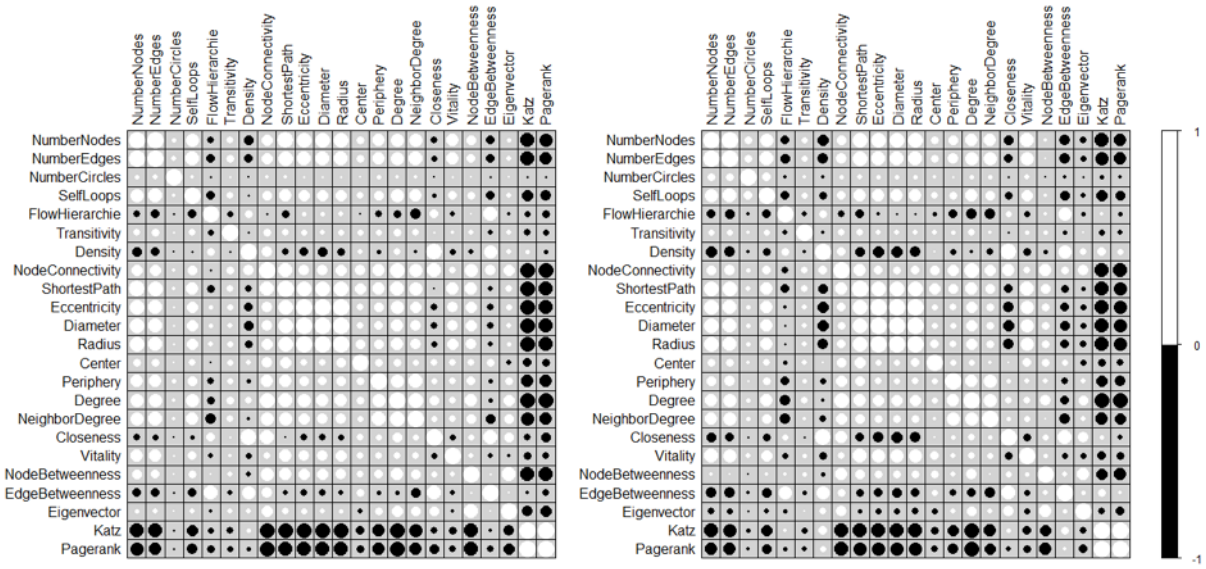


Figure 2.3: Correlation matrices for shop 1 (left) and shop 2 (right)

at the average value of page views for each visitor, users visiting the second shop look on average at more webpages per session compared to visitors of the first shop. Still, sessions of website visitors of the first shop result in more purchase conversions than in case of the second shop, where around 0.7 percent less purchases have taken place.

Table 2.4: Descriptive overview of our final datasets

	Shop 1	Shop 2
Users	58,545	18,759
Sessions	80,184	32,850
Purchase	(5.31%) 4,256	(4.63%) 1,520
Page views	692,975	475,500
Avg. page views	8.64	14.47

2.4.2 Correlation Analysis of Graph Measures

In the first step, we calculate the correlation matrix of the graph metrics to understand which features embody similar information about the navigational structure of a user's journey on a website. From each set of highly collinear variables, we select only one variable for further analysis to avoid issues of multicollinearity. The corresponding correlation matrices for both datasets are shown in Figure 2.3. We see that within the three graph metrics categories – structural, distance and centrality – high correlation exists between subsets of the variables. Partly, this is not a surprising result since some measures are either variations of each other (e.g., eigenvector, katz and pagerank centrality) or their calculation is based upon another metric (e.g. eccentricity and the related metrics diameter and radius).

For both shops, we observe similar correlation patterns. The measure *NumberNodes* is highly correlated with *NumberEdges*. Additionally, the three metrics *Eccentricity*, *Radius* and *Diameter* contain almost the same informational content. Furthermore, the centrality measures katz

and pagerank centrality – themselves highly correlated amongst each other - are negatively correlated with several graph metrics such as the structural components (number of nodes and edges) and the distance-based measures *ShortestPath*, *Eccentricity*, *Diameter* and *Radius*. The correlation of *NumberNodes* and *NumberEdges* can be interpreted in such a way that users tend to often perform as many click events as they visit unique webpages, i.e. webpages are generally visited only once and not several times by a user in a session. The correlation of *Eccentricity*, *Diameter* and *Radius* is unsurprising since they are based on the same basis, i.e. diameter being the maximum and radius being the minimum eccentricity in the graph.

To mitigate the issues of multicollinearity among the graph features, we remove highly correlated features on the basis of their variance inflation factor (VIF) (Alin, 2010) calculated as

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}, \tag{2.2}$$

where R^2 is the coefficient of determination from the regression of the covariate j on all other covariates (Stine, 1995). In contrast to the correlation coefficient, the VIF estimates the dependency of one covariate on all other covariates simultaneously, thus avoiding issues of the pairwise comparison. The higher the value of the VIF, the higher the correlation between the covariant j and all other variables. In general, covariates exceeding a VIF value between five and ten are seen as being prone to multicollinearity (Babin et al., 2005; Katrutsa & Strijov, 2017). We set our threshold to five and remove covariates exceeding this VIF value from the feature set. The calculation of the VIF is done in a step-wise manner, since the removal of a variable with high correlation affects the remaining variables influence. We recalculate the VIF for all remaining variables after removing the covariant with the highest VIF value from the preceding evaluation round. For both shops VIF results were almost consistent leading to the elimination of the covariates *Katz* (VIF = 2536.10 for shop 1 / 1201.07 for shop 2), *Diameter* (VIF = 176.74 / 758.24), *NumberNodes* (VIF = 166.33 / 200.10), *NodeConnectivity* (VIF = 82.06 / 45.88), *Closeness* (VIF = 30.37 / 22.65), *Pagerank* (VIF = 25.31 / 15.43), *ShortestPath* (VIF = 14.97 / 10.42), *Degree* (VIF = 10.59 / 5.08) and *NumberEdges* (VIF = 7.02 / 5.08) from the feature set. Additionally, in case of shop 1 the covariant *Eccentricity* (VIF = 993.34) and in case of shop 2 the *Radius* (VIF = 264.23) exceed the VIF threshold. Since these two metrics show high VIF values from the very beginning which drop significantly once either of the two is removed, we remove the covariant with the higher overall VIF, *Eccentricity,* (VIF = 993.34 for shop 1 / 257.25 for shop 2) for both shops and keep the covariant *Radius* in both datasets to increase consistency and facilitate the analysis.

This results in a final feature set consisting of 13 graph metrics, which we use for further analysis.

2.4.3 Predictive Performance

Using the subset of the 13 remaining graph features, we compare their predictive performance against the traditional feature set based on the GLM, RF and GBT algorithms introduced above.

Looking at AUC-PR (Table 2.5), we observe that the graph-based approach outperforms the traditional set of variables in all six instances independent of the underlying model. We further observe that the RF performs worse compared to GBT and the linear GLM for both shops. Furthermore, with the exception of the RF model, both models achieve higher AUC-PR values in case of shop 1. All models outperform the expected performance of a random model equal to the purchase rate of 5.3% and 4.6% for shop 1 and 2, respectively.

Table 2.5: AUC-PR values for shop 1 and shop 2 for the applied models

Model	GLM		RF		GBT	
Covariates	Graph	Traditional	Graph	Traditional	Graph	Traditional
Shop 1	0.372	0.271	0.287	0.262	0.372	0.262
Shop 2	0.311	0.243	0.300	0.247	0.317	0.288

For the lift measure, we observe that all three models trained on the applied graph metrics constitute for a clear improvement compared to random targeting. Figure 2.4 visualizes model lift in a gain chart for the three models on each dataset. Intuitively, the gain chart provides information about the number of purchasers if n% of users are targeted by the model, for example with a marketing incentive. Along the x-axis all views are plotted ordered by their predicted probability to purchase starting with those views having the highest probability. The y-axis represents the cumulative number of purchases among those page views. The upper grey bound shows the outcome of a perfect model which classifies all views according to their correct outcome, while a random model would result in a 45-degree diagonal. The steeper the curve is for a model, the better the model.

In case of the first shop, for the first around 30 percent of samples tested, both the GLM and the GBT model perform almost equally, i.e. their performance in terms of classifying views with a high predicted purchase probability. In the beginning, RF performs slightly worse until around 30 percent of the samples are tested where the model exhibits a similar performance compared to the other two applied models but is soon visibly outperformed for larger samples. For the second shop, all three models are even more homogenous in terms of their predictive performance until a threshold of around 50 percent of samples is reached. Exceeding this threshold we observe again a similar performance of GLM and GBT, while the RF model falls behind.

In general, the GLM model performs comparable in terms of lift compared to GBT. This is surprising considering the general performance of the models and the ability of GBT to model non-linear relations between the predictors. Given that all graph metrics are different measures to describe the same underlying graph structure, the good performance of the logistic regression model might be an indication that there are no significant non-linear dependencies between the graph metrics in predicting purchase behavior

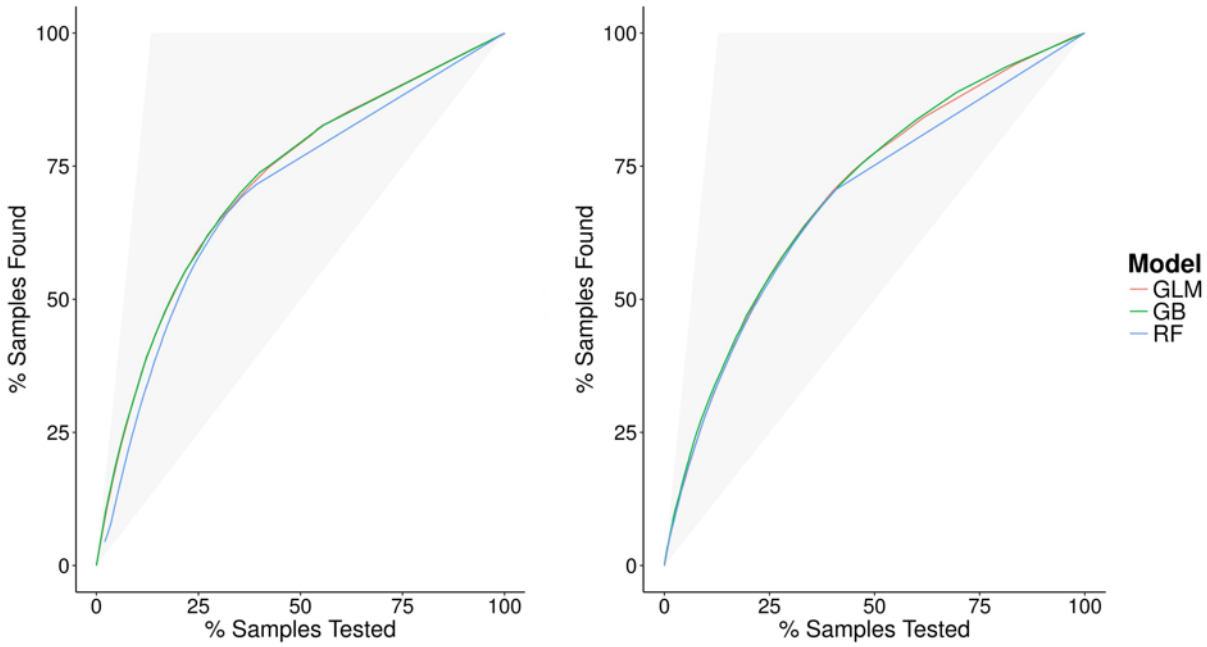


Figure 2.4: Lift chart for shop 1 (left) and shop 2 (right)

2.4.4 Variable Importance

In order to shed light on the direction of effect and performance of each graph metric, we analyze the model-wise importance of each graph measure. In case of the GLM model, we report the raw and exponentiated coefficients for each dataset. Due to the large data size, we observe that almost all coefficients are highly significant even at the 0.01% level. The coefficient of the variable *NumberCircles* is least significant at the 0.1% level. We thus focus on the analysis of effect size.

Since all variables are standardized, we analyze their effect size expressed in terms of the impact a change by one standard deviation (SD) has on the odds ratio (Table 2.6). With the odds ratio defined by the ratio of probabilities $P(Purchase)/P(No\ purchase)$, an exponentiated effect above 1 indicates a larger purchase probability. For both shops, *Radius* and *SelfLoops* have a strong positive effect on purchase probability. Specifically, an increase in *Radius* by one standard deviation, associated with less compact graphs, indicates an increase in purchase odds by 135% (shop 1) or 51% (shop 2), while an increase in *SelfLoops* by one SD leads to an increase in purchase odds by 40% (shop 1) or 51% (shop 2). A slightly smaller effect exists for *EdgeBetweenness* where a one SD increase, due to less connections between nodes, is associated with a 12% (shop 1) or 28% (shop 2) increase. In contrast, we observe the largest negative impact on purchase odds for a decrease in *Density*, where a decrease by one SD, observed for sparser graphs, increases the odds of a purchase by 23% ($1/0.81 = 1.23$) (shop 1) or 39% (shop 2). *FlowHierarchy* is estimated to have a negative effect of similar size. While there are some differences in effect size, we observe no difference in direction for the above variables, which have the strongest impact. In sum, the observed pattern suggests that linear click-paths related to search behavior may be more indicative of users with purchase intention.

The variables *NumberCircles, Eigenvector NeighborDegree,* and *Center* show coefficients in different directions between shop 1 and 2, indicating that the underlying relationship may be shop dependent to a larger degree.

Table 2.6: Estimated coefficients for the GLM model

	GLM Model for Shop 1			GLM Model for Shop 2		
Variable	Coefficient	Std. Error	Odds Ratio	Coefficient	Std.Error	Odds Ratio
Intercept	-0.29 ***	0.004	0.75	-0.29 ***	0.005	0.75
NumberCircles	-0.04 **	0.014	0.96	0.09 ***	0.013	1.09
Density	-0.21 ***	0.006	0.81	-0.33 ***	0.009	0.72
Vitality	-0.12 ***	0.006	0.89	-0.06 ***	0.009	0.94
NodeBetweenness	0.05 ***	0.005	1.05	0.03 ***	0.006	1.03
EdgeBetweenness	0.11 ***	0.008	1.12	0.25 ***	0.010	1.28
Eigenvector	-0.06 ***	0.005	0.94	0.02 **	0.006	1.02
Radius	0.86 ***	0.007	2.35	0.41 ***	0.008	1.51
SelfLoops	0.34 ***	0.005	1.40	0.51 ***	0.007	1.66
FlowHierarchy	-0.20 ***	0.006	0.82	-0.25 ***	0.007	0.78
NeighborDegree	-0.15 ***	0.007	0.86	0.05 ***	0.006	1.06
Center	-0.10 ***	0.005	0.91	0.03 ***	0.006	1.03
Periphery	0.14 ***	0.005	1.15	0.05 ***	0.005	1.05
Transitivity	0.08 ***	0.004	1.09	0.11 ***	0.005	1.11

Significance levels: 0.0001 '***' 0.001 '**' 0.01 '*'

For the gradient boosted trees, we calculate the variable bag importance for both datasets based on the weighted increase in node purity for the splits on each variable averaged over all trees (Hastie et al., 2009). In other words, the variable importance captures the relative contribution to improve classification for each variable in the model. The variable importance scores are scaled to sum up to 100 and are reported in Figure 2.5. The importance ranking for the non-linear GBT model shows different patterns compared to the logit coefficient analysis in so far as *Radius, Density* and *FlowHierarchy* are only marginally relevant for purchase prediction while *Vitality* and *SelfLoops* constitute the most important variables for both shops. However, we observe that feature importance is centered around *Vitality* with a sharp decrease towards the *SelfLoops*, the second most important variable. Since high values of *Vitality* refer to the existence of important connections in the graph structure, the importance of the variable could be explained through being able to detect specific user behavior, i.e. signifying either goal-oriented, non-recursive browsing behavior or the existence of bridging elements in the user website journey such as overview or search result pages.

Additionally, there exists some deviation in variable importance between shops. In case of shop 1 the variables *Radius* and *Density* are the most important among the remaining variables, whereas in case of shop 2 this is true for the variables *NeighborDegree* and *Transitivity*. While *Radius* and *Density* refer to global characteristics of the clickstream graph, *NeighborDegree* and *Transitivity* are related to direct neighborhoods of single nodes. These two feature pairs could flag different browsing behaviors present in both shops. All other variables only constitute for a small percentage in terms of variable importance and seem to be negligible for the distinction of purchasers and non-purchasers in case of the two datasets applied and the GBT model.

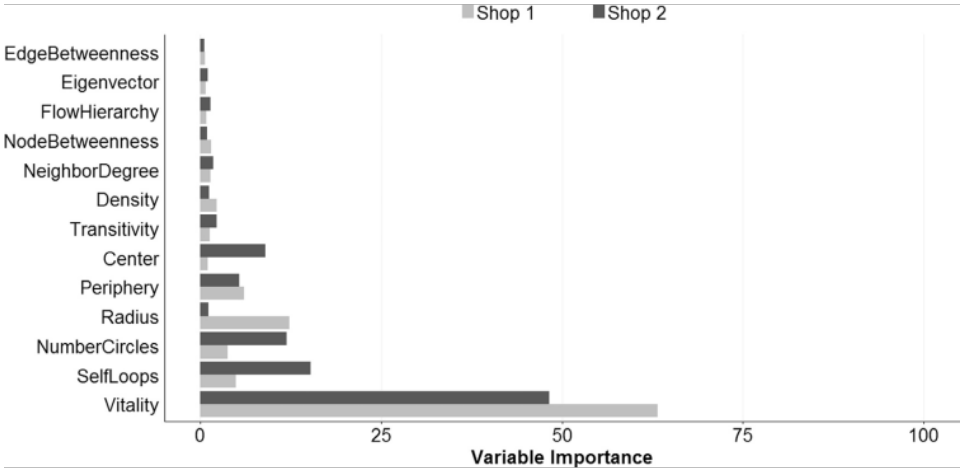


Figure 2.5: Variable importance for the gradient boosting model for shop 1 and 2

We use Partial Dependence Plots (PDP) for a deeper analysis of non-linear effects of each variable as estimated by the GBT model (Hastie et al., 2009). PDP are a graphical tool to examine the marginal effect of each variable on the model prediction accounting for the (average) effects of all other variables. Figure 2.6 and Figure 2.7 show the PDP for shop 1 and shop 2, respectively. For both shops, we observe similar patterns supporting the between-shops robustness of effects also observed for GLM. Naturally, the PDP for both shops show the most distinctive patterns for the variables with the highest important scores (see Figure 2.5).

According to the PDP an increasing value of *Vitality*, being the most important variable for both shops and which we interpret as a rising number of central pages in the user journey, is linked to an increase in purchase probability for both shops. Given that a rather high value of *Vitality* represents a linear, non-recursive type of browsing behavior, this evidence might indicate that the feature is especially relevant for detecting users who show such kind of behavior. Furthermore, since *Vitality* represents the change of distances for all present nodes in the graph, this metric might additionally be able to capture users with a high number of page views which might be an indication of browsing behavior leading to a purchase. The PDP of *SelfLoops*, which captures re-occurring webpage visits and constitutes the second most important variable for both shops, reveals a similar link. For both shops the purchase probability increases with the number of times a user revisits the same page

In case of shop 1, the lower the value of *Density* and the higher the *Radius*, the more likely a purchase occurs in case of shop 1, which again signals the goal-oriented shopping behavior of users visiting one page after another. In case of shop 2, the variable importance scores of *NeighborDegree* and Transitivity have been shown to be relevant for predicting purchasers as captured by the features important scores. The PDP for *NeighborDegree* illustrates the relationship that the higher the value of this metrics, signaling extensive browsing behavior within a direct neighborhood of a webpage, the more likely it is that a purchase occurs. For *Transitivity* no clear relationship is observable based on the PDP. However, for *NumberCircles* and *Periphery* a sharp jump in the slope indicates that there is a strong increase in purchase

probability after reaching a certain threshold.

Altogether these might be indicators that both shops constitute for different shopping behavior. Whereas the first shop reflects rather goal-oriented behavior, the second represents a browsing-related shopping experience. However, given the predictive value of the graph metrics, in-depth analysis beyond the scope of this paper will be necessary to identify the specific user intentions associated with a graph structure and could focus on experimental investigation of the link between stated user intention and each metric and establishing the robustness of the observed dependencies structures to different shops and product categories.

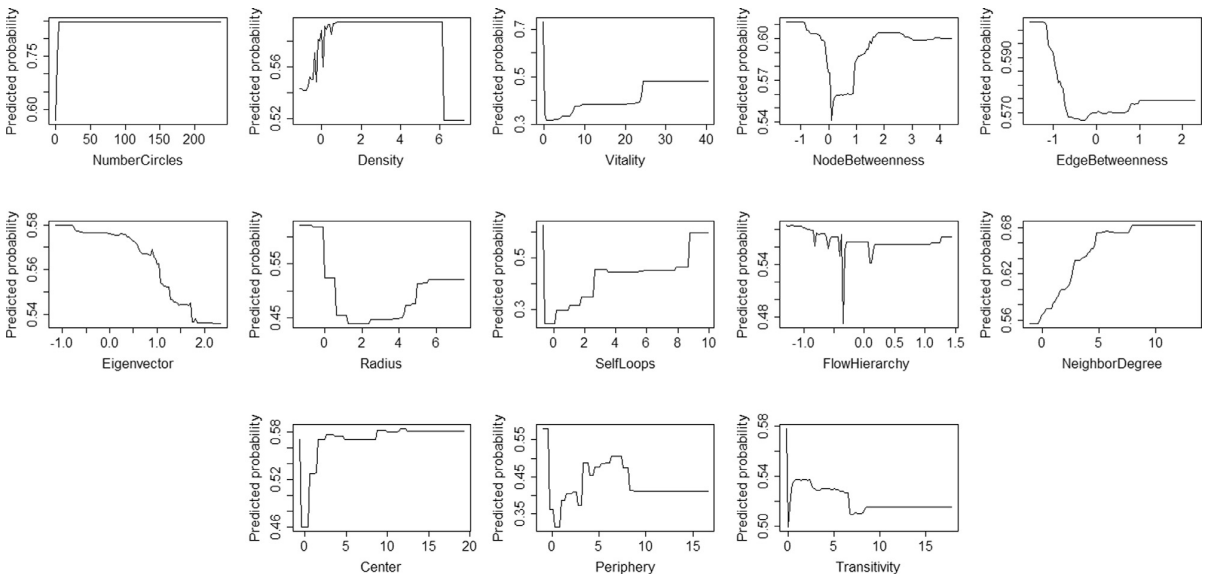


Figure 2.6: Partial dependence plots for shop 1

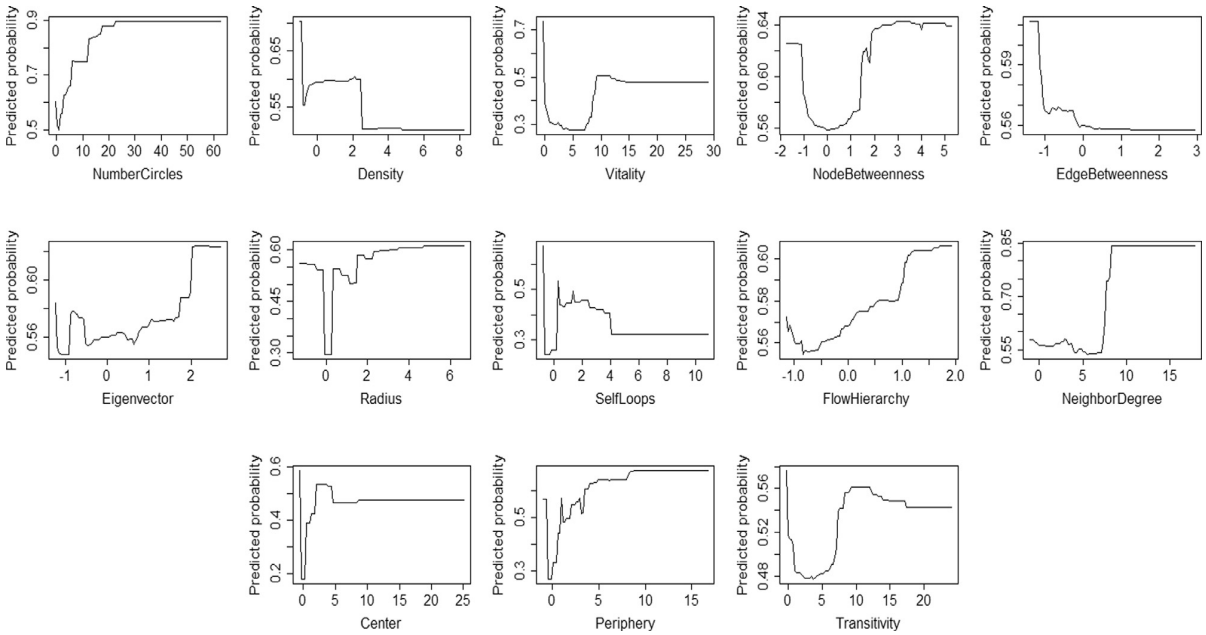


Figure 2.7: Partial dependence plots for shop 2

2.5 Conclusion

Using real-life clickstream datasets of two different shops we observe for both the linear GLM model and the non-linear Random Forest model that distance- and centrality-based graph metrics are effective in predicting purchase behavior of users. We derived user-centered, session-based graphs from clickstream data, where each graph is developed incrementally, i.e. each new page view of the user develops the graph further. Each of the 23 tested graph metrics are calculated for each intermediate state of a graph. We report and control for multicollinearity between the graph metrics by pre-processing using variable inflation factors and train three selected high-performing algorithms on the resulting dataset. Independent of the employed model, the proposed variables result in a substantial increase in the area-under-the-precision-recall-curve and model lift in predictive power compared to random targeting and a set of standard aggregation features derived from clickstream.

Looking at the importance of each graph metric, we observe clear differences in the relevance of variables between the linear and non-linear models. We suggest that closeness vitality in particular followed by closeness centrality and the number of self-loops should be considered promising candidates in future applications.

We also identify some promising areas for future research. An alternative approach to calculate graph metrics could include different graph construction methods such as using bi-partite graphs, where two different types of nodes are included, to represent the structure of a user session in more detail. Additionally, constructing weighted graphs by rating frequently taken paths as more important or taking into account the time spent on specific pages could improve the representation of the users' journey on a website and consequently increase the accuracy when predicting outcome of a session.

Bibliography

Alin, A. (2010). Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(3), 370–374. https://doi.org/10.1002/wics.84

Anitha, A. (2010). A new web usage mining approach for next page access prediction. *International Journal of Computer Applications*, *8*(11), 7–10. https://doi.org/10.5120/1252-1700

Antonellis, P., Makris, C., & Tsirakis, N. (2009). Algorithms for clustering clickstream data. *Information Processing Letters*, *109*(8), 381–385. https://doi.org/10.1016/j.ipl.2008.12.011

Babin, B., Anderson, R. E., Hair, J. F., & Black, B. (2005). *Multivariate Data Analysis* (Sixth). TBS.

Banerjee, A., & Ghosh, J. (1997). Clickstream Clustering using Weighted Longest Common Subsequences, In *Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining*, Chicago, IL, USA.

Berka, P., & Labský, M. (2007). Predicting Page Occurrence in a Click-Stream Data: Statistical and Rule-Based Approach. In P. Perner (Ed.), *Advances in Data Mining: Theoretical Aspects and Applications* (pp. 135–147). Berlin, Heidelberg, Springer. https://doi.org/10.1007/978-3-540-73435-2_11

Berry, M. J. A., & Linoff, G. (2011). *Data Mining Techniques: For Marketing, Sales and Customer Relationship Management* (Third). Indianapolis, Wiley.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Byeon, H. (2013). Evaluating the online buying behavior using network analysis. *International Journal of Advancements in Computing Technology*, *5*(12).

Chan, T., I, J., Macasaet, C., Kang, D., Hardy, R. M., Ruiz, C., Porras, R., Baron, B., Qazi, K., Padraic, H., & Honda, T. (2014). Predictive Models for Determining If and When to Display Online Lead Forms, In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, Québec City, Québec, Canada, AAAI Press.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Colladon, A. F., & Remondi, E. (2017). Using social network analysis to prevent money laundering. *Expert Systems with Applications*, *67*, 49–58. https://doi.org/http://dx.doi.org/10.1016/j.eswa.2016.09.029

eMarketer. (2019a). Global E-Retail Growth Rate 2023.

eMarketer. (2019b). Global Retail E-Commerce Market Size 2014-2023.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/http://dx.doi.org/10.1016/j.patrec.2005.10.010

Fitzpatrick, T., & Mues, C. (2016). An empirical comparison of classification algorithms for mortgage default prediction: Evidence from a distressed mortgage market. *European Journal of Operational Research*, *249*(2), 427–439. https://doi.org/http://dx.doi.org/10.1016/j.ejor.2015.09.014

Girija, P., & Kavitha, V. (2013). An approach for predicting user's web access pattern. *International Journal of Computer Science and Management Research*, *2*(5), 2585–2589.

Gündüz, Ş., & Özsu, M. T. (2003). A Web Page Prediction Model Based on Click-Stream Tree Representation of User Behavior, In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '03)*, Washington, D.C., ACM Press. https://doi.org/10.1145/956750.956815

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The Elements of Statistical Learning* (Second). New York, Springer.

He, Y.-l., Liu, J. N. K., Hu, Y.-x., & Wang, X.-z. (2015). OWA operator based link prediction ensemble for social network. *Expert Systems with Applications*, *42*(1), 21–50. https://doi.org/http://dx.doi.org/10.1016/j.eswa.2014.07.018

Hong, T., & Kim, E. (2012). Segmenting customers in online stores based on factors that affect the customer's intention to purchase. *Expert Systems with Applications*, *39*(2), 2127–2131. https://doi.org/10.1016/j.eswa.2011.07.114

Iwanaga, J., Nishimura, N., Sukegawa, N., & Takano, Y. (2016). Estimating product-choice probabilities from recency and frequency of page views. *Knowledge-Based Systems*, *99*, 157–167. https://doi.org/10.1016/j.knosys.2016.02.006

Jiang, Q., Tan, C.-H., & Wei, K.-K. (2012). Cross-Website Navigation Behavior And Purchase Commitment: A Pluralistic Field Research, In *Proceedings of the Pacific Asia Conference on Information Systems (PACIS'12)*, AIS.

Kalczynski, P. J., Senecal, S., & Nantel, J. (2006). Predicting on-line task completion with clickstream complexity measures: A graph-based approach. *International Journal of Electronic Commerce*, *10*(3), 121–141. https://doi.org/10.2753/jec1086-4415100305

Katrutsa, A., & Strijov, V. (2017). Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. *Expert Systems with Applications*, *76*, 1–11. https://doi.org/10.1016/j.eswa.2017.01.048

Kitts, B., Hetherington-Young, K., & Vrieze, M. (2002). Large-scale mining, discovery and visualization of user clickpaths. *International Journal of Image and Graphics*, *02*(01), 21–48. https://doi.org/10.1142/S0219467802000536

Lee, H., Choi, S. Y., & Kang, Y. S. (2009). Formation of e-satisfaction and repurchase intention: Moderating roles of computer self-efficacy and computer anxiety. *Expert Systems with Applications*, *36*(4), 7848–7859. https://doi.org/10.1016/j.eswa.2008.11.005

Lee, M., Ferguson, M. E., Garrow, L. E., & Post, D. (2010). *The Impact of Leisure Travelers' Online Search and Purchase Behavior on Promotion Effectiveness.*

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: Online Appendix. *European Journal of Operational Research*, *247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Ling, C. X., & Li, C. (1998). Data Mining for Direct Marketing: Problems and Solutions, In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, Menlo Park, US, AAAI Press.

Lu, L., Dunham, M., & Meng, Y. (2005). Mining Significant Usage Patterns from Clickstream Data, In *Proceedings of the 7th International Conference on Knowledge Discovery on the Web*, Chicago, Springer. https://doi.org/10.1007/11891321_1

Martens, D., Provost, F., Clark, J., & Junqué de Fortuny, E. (2016). Mining massive fine-grained behavior data to improve predictive analytics. *MIS Quarterly*, *40*(4), 869–888. https://doi.org/10.25300/misq/2016/40.4.04

Moe, W. W. (2003). Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of Consumer Psychology*, *13*(1), 29–39. https://doi.org/10.1207/S15327663JCP13-1\&2_03

Moe, W. W., Chipman, H., George, E. I., & McCulloch, R. E. (2002). *A Bayesian Tree Model of Online Purchasing Behavior Using In-Store Navigational Clickstream.*

Moe, W. W., & Fader, P. S. (2004). Capturing evolving visit behavior in clickstream data. *Journal of Interactive Marketing*, *18*(1), 5–19. https://doi.org/10.1002/dir.10074

Montgomery, A. L., Li, S., Srinivasan, K., & Liechty, J. C. (2004). Modeling online browsing and path analysis using clickstream data. *Marketing Science*, *23*(4), 579–595. https://doi.org/10.1287/mksc.1040.0073

Óskarsdóttir, M., Bravo, C., Verbeke, W., Sarraute, C., Baesens, B., & Vanthienen, J. (2017). Social network analytics for churn prediction in telco: Model building, evaluation and network architecture. *Expert Systems with Applications*, *85*, 204–220. https://doi.org/10.1016/j.eswa.2017.05.028

Padmanabhan, B., Zhiqiang, Z., & Kimbrough, S. O. (2006). An empirical analysis of the value of complete information for ECRM models. *MIS Quarterly*, *30*(2), 247–267.

Pai, D., Sharang, A., Yadagiri, M. M., & Agrawal, S. (2014). Modelling Visit Similarity Using Click-Stream Data: A Supervised Approach, In *International Conference on Web Information Systems Engineering*, Cham, Springer.

Panagiotelis, A., Smith, M. S., & Danaher, P. J. (2014). From Amazon to Apple: Modeling online retail sales, purchase incidence, and visit behavior. *Journal of Business & Economic Statistics*, *32*(1), 14–29. https://doi.org/10.1080/07350015.2013.835729

Park, C. H., & Park, Y.-H. (2016). Investigating purchase conversion by uncovering online visit patterns. *Marketing Science*, *35*(6), 894–914. https://doi.org/10.1287/mksc.2016.0990

Park, S., Suresh, N. C., & Jeong, B.-K. (2008). Sequence-based clustering for Web usage mining: A new experimental framework and ANN-enhanced K-means algorithm. *Data & Knowledge Engineering*, *65*(3), 512–543. https://doi.org/10.1016/j.datak.2008.01.002

Piatetsky-Shapiro, G., & Masand, B. (1999). Estimating Campaign Benefits and Modeling Lift, In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM. https://doi.org/10.1145/312129.312225

Pitman, A., & Zanker, M. (2010). Insights from Applying Sequential Pattern Mining to E-commerce Click Stream Data, In *2010 IEEE International Conference on Data Mining Workshops*. https://doi.org/10.1109/ICDMW.2010.31

Sarwar, S. M., Hasan, M., & Ignatov, D. I. (2015). Two-stage Cascaded Classifier for Purchase Prediction. *arXiv preprint*, arXiv:1508.03856 [cs].

Sato, S., & Asahi, Y. (2012). A daily-level purchasing model at an e-commerce site. *International Journal of Electrical and Computer Engineering (IJECE)*, *2*(6), 831–839. https://doi.org/10.11591/ijece.v2i6.1816

Schult, D. A. (2008). Exploring Network Structure, Dynamics, and Function Using NetworkX, In *Proceedings of the 7th Python in Science Conference*.

Senecal, S., Kalczynski, P. J., & Fredette, M. (2014). Dynamic identification of anonymous consumers' visit goals using clickstream. *International Journal of Electronic Business*, *11*(3), 220. https://doi.org/10.1504/IJEB.2014.063036

Shams, B., & Haratizadeh, S. (2017). Graph-based collaborative ranking. *Expert Systems with Applications*, *67*, 59–70. https://doi.org/10.1016/j.eswa.2016.09.013

Sismeiro, C., & Bucklin, R. E. (2004). Modeling purchase behavior at an e-commerce web site: A task-completion approach. *Journal of Marketing Research*, *41*(3), 306–323. https://doi.org/10.1509/jmkr.41.3.306.35985

Stange, M., & Funk, B. (2015). How Much Tracking Is Necessary? - The Learning Curve in Bayesian User Journey Analysis, In *Proceedings of the European Conference on Information Systems Completed Research Papers*.

Stine, R. A. (1995). Graphical interpretation of variance inflation factors. *The American Statistician*, *49*(1), 53–56. https://doi.org/10.1080/00031305.1995.10476113

Suh, E., Lim, S., Hwang, H., & Kim, S. (2004). A prediction model for the purchase probability of anonymous customers to support real time web marketing: A case study. *Expert Systems with Applications*, *27*(2), 245–255.

Takaya Saito, & Marc Rehmsmeier. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*, *10*(3), e011843.

Van den Poel, D., & Buckinx, W. (2005). Predicting online-purchasing behaviour. *European Journal of Operational Research*, *166*(2), 557–575. https://doi.org/10.1016/j.ejor.2004.04.022

VanderMeer, D., Dutta, K., Datta, A., Ramamritham, K., & Navanthe, S. B. (2000). Enabling Scalable Online Personalization on the Web, In *Proceedings of the 2nd ACM Conference on Electronic Commerce (EC'00)*, Minneapolis, Minnesota, United States, ACM Press. https://doi.org/10.1145/352871.352892

Vroomen, B., Donkers, B., Verhoef, P. C., & Franses, P. H. (2005). Selecting profitable customers for complex services on the internet. *Journal of Service Research*, *8*(1), 37–47. https://doi.org/10.1177/1094670505276681

Wu, F., Chiu, I.-H., & Lin, J.-R. (2005). Prediction of the Intention of Purchase of the User Surfing on the Web Using Hidden Markov Model, In *Proceedings of the 2005 International Conference on Services Systems and Services Management (ICSSSM'05)*, IEEE. https://doi.org/10.1109/ICSSSM.2005.1499501

Zhang, Y., Bradlow, E. T., & Small, D. S. (2015). Predicting customer value using clumpiness: From RFM to RFMC. *Marketing Science*, *34*(2), 195–208. https://doi.org/10.1287/mksc.2014.0873

Zhao, Y., Yao, L., & Zhang, Y. (2016). Purchase prediction using Tmall-specific features: Purchase prediction using Tmall-specific features. *Concurrency and Computation: Practice and Experience*, *28*(14), 3879–3894. https://doi.org/10.1002/cpe.3720

Zheng, Z., Padmanabhan, B., & Kimbrough, S. O. (2003). On the existence and significance of data preprocessing biases in web-usage mining. *Journal on Computing*, *15*(2), 148. https://doi.org/10.1287/ijoc.15.2.148.14449

2.A Appendix

Table 2.7: Summary statistics of the graph metrics for each shop

Shop 1							
	Min.	**25%Q.**	**Median**	**Mean**	**75% Q.**	**Max.**	**Std. Dev.**
Purchase	0.00			0.14		1.00	
NumberCircles	0.00	0.00	1.00	3.36	3.00	38150.00	132.95
Density	0.00	0.13	0.25	0.30	0.42	2.00	0.26
Vitality	0.00	1.00	7.50	54.51	36.00	25818.68	289.35
NodeBetweenness	0.00	0.00	0.15	0.12	0.17	0.50	0.09
EdgeBetweenness	0.00	0.13	0.20	0.21	0.28	0.50	0.14
Eigenvector	0.00	0.00	0.21	0.19	0.33	0.58	0.18
Radius	0.00	1.00	2.00	1.79	2.00	24.00	1.38
SelfLoops	0.00	0.00	0.00	0.67	1.00	13.00	1.02
FlowHierarchy	0.00	0.13	0.44	0.49	1.00	1.00	0.38
NeighborDegree	0.00	0.33	1.00	1.11	1.63	23.91	1.09
Center	1.00	1.00	1.00	1.58	2.00	20.00	0.85
Periphery	1.00	2.00	2.00	2.49	3.00	30.00	1.49
Transitivity	0.00	0.00	0.00	0.01	0.00	1.00	0.05
Shop 2							
	Min.	**25%Q.**	**Median**	**Mean**	**75%Q.**	**Max.**	**Std. Dev.**
Purchase	0.00			0.13		1.00	
NumberCircles	0.00	1.00	2.00	7.49	7.00	14298.00	117.34
Density	0.00	0.09	0.18	0.25	0.33	2.00	0.23
Vitality	0.00	4.00	24.29	225.33	125.17	29824.62	897.00
NodeBetweenness	0.00	0.09	0.14	0.12	0.17	0.50	0.08
EdgeBetweenness	0.00	0.09	0.15	0.18	0.22	0.50	0.12
Eigenvector	0.00	0.00	0.18	0.19	0.32	0.58	0.16
Radius	0.00	1.00	2.00	2.76	4.00	28.00	2.32
SelfLoops	0.00	0.00	1.00	1.09	1.00	37.00	1.64
FlowHierarchy	0.00	0.12	0.33	0.42	0.67	1.00	0.35
NeighborDegree	0.00	0.80	1.39	1.79	2.23	48.61	2.03
Center	1.00	1.00	1.00	1.62	2.00	31.00	1.03
Periphery	1.00	2.00	2.00	2.89	3.00	39.00	1.98
Transitivity	0.00	0.00	0.00	0.01	0.00	1.00	0.04

Chapter 3

Targeting Customers for Profit: An Ensemble Learning Framework to Support Marketing Decision-Making

Publication
Lessmann, S., Haupt, J., Coussement, K., & De Bock, K. W. (2019). Targeting customers for profit: An ensemble learning framework to support marketing decision making. Information Sciences, In Press. https://doi.org/10.1016/j.ins.2019.05.027

Abstract
Marketing messages are most effective if they reach the right customers. Deciding which customers to contact is an important task in campaign planning. The paper focuses on empirical targeting models. We argue that common practices to develop such models do not account sufficiently for business goals. To remedy this, we propose profit-conscious ensemble selection, a modeling framework that integrates statistical learning principles and business objectives in the form of campaign profit maximization. Studying the interplay between data-driven learning methods and their business value in real-world application contexts, the paper contributes to the emerging field of profit analytics and provides original insights how to implement profit analytics in marketing. The paper also estimates the degree to which profit-conscious modeling adds to the bottom line. The results of a comprehensive empirical study confirm the business value of the proposed ensemble learning framework in that it recommends substantially more profitable target groups than several benchmarks.

3.1 Introduction

Business analytics revolutionizes the face of decision support. Skepticism toward formal decision aids used to be widespread among executives. Today, we witness an unprecedented interest in quantitative decision aids and analytic models. Vast amounts of data, powerful pattern extraction algorithms, and easy to use software systems fuel this development and promise to improve management support. The paper concentrates on decision support in marketing campaign planning. Campaign planners need to answer three questions (Elsner et al., 2004): when to make an offer (timing), how often to make an offer (frequency), and whom to contact (target group selection). We focus on the target group selection problem, which has been studied in the direct marketing and churn management literature (Zhu et al., 2017). To target

marketing offers, companies use response models, which estimate acceptance probabilities for individual customers. Corresponding predictions facilitate targeting the most likely responders.

Modeling response behavior on the level of an individual customer is a popular use case of business analytics in marketing. Developments in the scope of big data have a sizeable impact on customer response modeling, which we discuss along the well-known four V's volume, variety, velocity, and value that characterize big data. First, the volume dimension implies that companies have more detailed records of past customer behavior and information related to customer preferences (Martens et al., 2016). Such behavioral information enters response models in the form of novel attributes from which acceptance probability predictions are eventually derived. Second, the variety dimension refers to new and often unstructured sources of data, which companies can unlock for gaining business insight. The use of text analytics to extract information from product reviews, postings in social media, etc. illustrates this approach and contributes attitudinal information, which further expands to scope of customer characteristics that enter response models. Third, the velocity dimension postulates that novel data arrives with higher speed and implies a necessity to reduce the latency of decision-making. For example, response model-based targeting decisions in digital advertisement must be made in real-time and the number of application settings that also require real-time decision-making tends to increase in the big data era. Finally, there is much evidence of big data creating considerable value for marketing, which emerges from enhanced decision-making (Tambe, 2014).

Response models use a variety of prediction methods including, artificial neural networks, support vector machines, or tree-based approaches. However, prediction methods are designed for generality and support decision-making in many fields such as credit scoring (Maldonado, Pérez, et al., 2017) and fraud detection (Van Vlasselaer et al., 2017). Developing a prediction model involves minimizing a statistical loss function on a labeled training sample (Hastie et al., 2009). We argue that using an off-the-shelf method for customer targeting suffers a limitation. Contextual information related to the actual decision task does not enter model development. Budget constraints, customer lifetime value, parallel campaigns – relevant information in campaign planning – have no effect on the estimation of the targeting model. Therefore, the objective of the paper is to develop and test a contextualized modeling framework that accounts for business objectives during model development.

Current trends in marketing support this objective. Big data facilitates an increasing degree of personalization in marketing communication (Golrezaei et al., 2014). Likewise, an increasing amount of information is distributed through digital channels (Ding et al., 2015). These developments amplify the scale of targeting decisions and require decision-making in real-time. Therefore, marketers need to automate targeting decisions. A high recognition of business goals during model development seems especially important when targeting models operate in a self-governed manner. More generally, our focus on the business value of empirical decision support models echoes the recent call for a higher recognition of managerial objectives in modeling, which gave rise to the emerging field of profit analytics (Maldonado, Bravo, et al., 2017).

The contribution of the paper to the literature is threefold. First, we propose a new model-

ing methodology for profit-conscious ensemble selection (PCES). We design PCES in such a way that it integrates established principles of statistical inference with marketing objectives in customer targeting. A related design goal is to mimic the way in managers contextualize recommendation from model-based decision aids (Fuller & Dennis, 2009). PCES-based targeting models are contextualized in the sense that they account for marketing objectives and constraints at earlier stages of the model development process than existing approaches. We hypothesize that a contextualization of the model development process improves the quality of targeting decisions.

The second contribution stems from a comprehensive empirical analysis, which includes twenty-five real-world marketing data sets from different industries, of the effectiveness of alternative paradigms toward customer targeting. Beyond comparing an arsenal of alternative targeting models, we contrast three fundamentally different modeling philosophies. The first approach, which we refer to as *profit-agnostic,* relies on statistical learning and develops targeting models through minimizing a statistical loss-function (Hastie et al., 2009). We consider this approach to represent standard practice in predictive analytics. The second approach derives targeting models from maximizing business performance while disregarding statistical learning principles. We consider this approach an extreme form of profit analytics and call corresponding models *profit-centered.* The third approach represents a hybrid solution in the form of PCES, which balances between statistical and economic considerations. This three-facetted empirical design provides novel insight concerning the relative merits of fundamentally different approaches toward predictive modeling.

The empirical design also facilitates the third and last contribution of the paper. In particular, the paper provides an estimate of the degree to which incorporating business goals into prediction model development raises the business performance (e.g., return on marketing) of model-based (targeting) decisions. We achieve this through estimating the campaign profit that emerges from model-based targeting and the marginal profit of PCES-based targeting, respectively. Corresponding results provides a clear and managerially meaningful measure of the business value of a targeting model and the extent to which PCES improves decision quality.

3.2 Background and Related Work

Related work splits into three streams. First, prior work on decision support systems (DSS) provide theoretical foundations (*Stream 1*). Second, related studies in forecasting and machine learning consider the interplay between predictive models and their value implications in economic contexts but differ in the methodology they employ and applications they consider (*Stream 2*). We sketch the connections and differences to these streams in the following. Subsequently, we discuss previous research on marketing decision support and customer targeting (*Stream 3*), which is particularly related to this study.

Papers from *Stream 1* examine the antecedents of (model-based) DSS effectiveness and highlight the importance of a DSS exhibiting high fit for the decision task. However, managers can mitigate a lack of fit if given an opportunity to post-process DSS recommendations (Fuller &

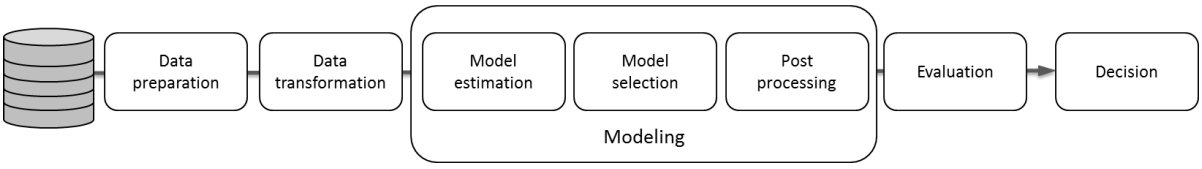


Figure 3.1: Simplified process of prediction model development without feedback loops between stages

Dennis, 2009). Specifically, managers' decision-making is guided by a mental model that enables them to appraise DSS outputs in awareness of a specific problem context, connect DSS outputs to decision quality, and, in this way, correct for misleading information from an inadequate model (Fuller & Dennis, 2009). This theory shows the merit of human supervision in model-based decision support and provides a design goal for the PCES approach developed here. We strive to combine the efficiency of automated, model-based decision-making with the ability of managers to improve decision quality through using contextual, task-specific information.

Prior work in *Stream 2* examines whether and when the development of data-driven prediction methods should account for economic objectives. Granger (1969) was the first to criticize the use of quadratic loss functions in forecasting and to propose loss functions that penalize positive and negative residuals differently. Subsequent studies contribute further theoretical insights and empirical evidence concerning asymmetric loss functions in forecasting (e.g. Christoffersen & Diebold, 1997). The cost-sensitive learning literature also studies asymmetric cost of error functions but focuses on classification models (Zhao & Li, 2017).

Research from *Stream 2* inspires the proposed PCES approach. PCES also employs non-standard, asymmetric loss functions for the development and assessment of predictive models. The main differences lie in the methodology and application. We focus on multivariate machine learning models as opposed to univariate time series models in forecasting. Our focus on decision problems in marketing campaign planning also implies that we study a different business objective (i.e., campaign profit). Specifically, the different errors in campaign planning are soliciting customers who do not respond and failing to contact customers who would respond (e.g., purchase an item) otherwise. This perspective on model errors is similar to cost-sensitive learning. Cost-sensitive learning, however, aims at generality. While generality is a goal worth pursuing, a DSS approach that focuses on a specific application context better reflects the unique characteristics and requirements of this context (Lilien, 2011). PCES is such an approach for decisions in the scope of targeted marketing where campaigns typically solicit only a small fraction of responsive customers. This implies a different notion of model performance compared to cost-sensitive learners, the objective of which is to minimize overall error costs (Zhao & Li, 2017).

Finally, there is a large body of literature on predictive models for customer targeting (*Stream 3*). Previous work has studied all steps of the predictive modeling process, which we depict in Figure 3.1. In interpreting Figure 3.1, it is important to note that we deliberately refrain from incorporating feedback loops. Research on data preparation includes endeavors to build an analytic database from past campaigns and test mailings (Rokach et al., 2008). Marketing

papers in the field data preparation examine how alternative definitions of the modeling target (e.g. Glady et al., 2009) or covariates (Mitrović et al., 2018) affect model quality. The data transformation step has been studied through the lens of feature selection (Maldonado, Bravo, et al., 2017) and independent variable projection (Coussement et al., 2017). The estimation of the actual marketing decision model, its tuning, and possible combination with other models (i.e., ensembling) is the process step that has attracted the largest attention in prior literature (Martens et al., 2016) and is also the focus of this paper. Other papers study a post-processing of model prediction to enhance calibration (e.g. Coussement & Buckinx, 2011) or design new indicators to measure the performance of a decision model (Verbraken et al., 2012).

The majority of previous studies estimate the targeting model using standard prediction methods (neural networks, support vector machines, etc.). We call this approach profit-agnostic because it does not take account of the actual decision context (i.e., customer targeting) and business objective (i.e., profit maximization) during model development. Only a few studies emphasize the inability of statistical accuracy indicators to reflect marketing objectives and propose application specific alternatives such as the (expected) maximum profit criterion for churn modeling (Verbeke et al., 2012; Verbraken et al., 2012). We add to this research through using a more general profit function, which enables us to study a broad range of targeting applications beyond churn. Focusing on profit-oriented model development, we also introduce the business goal earlier in the modeling process where corresponding information can exert more influence on the eventual model. To confirm this, we empirically compare PCES to the approach proposed in Verbeke et al. (2012).

To our knowledge, three studies consider a profit-oriented model development in marketing. Using a genetic algorithm (GA), Bhattacharyya (1999) estimates the parameters of a linear model so as to maximize profit. Stripling et al. (2018) further extends this approach to maximize the expected maximum profit criterion for churn modeling, while Cui et al. (2015) select customers with heterogeneous expected returns via partial ordering. PCES differs from these approaches in that it i) uses a more advanced ensemble learning paradigm and ii) adopts a multi-stage approach to balance statistical loss and business goals. To verify the appropriateness of this design, we empirically compare PCES to the GA-based approach of Bhattacharyya (1999) and Stripling et al. (2018).

Finally, research in information retrieval is concerned with ranking algorithms, for example to identify the top N most relevant search results for a query. Advanced solutions use deep learning in the form of convolutional neural networks to optimize ranking functions directly (Geng et al., 2016). Allocating marketing budgets in campaign planning could be framed as a ranking problem, so that corresponding advancements could have much potential to perform profit analytics in fundamentally new ways.[1]

[1]The authors would like to thank an anonymous reviewer for suggesting this approach toward profit analytics.

3.3 Methodology

In the following, we elaborate on our methodology. First, we review the statistical fundamentals of predictive models and explain how standard loss functions disregard application characteristics. Next, we discuss business goals in campaign planning and corresponding objective functions. Last, we elaborate on the PCES framework, which we propose to combine statistical and business objectives.

3.3.1 Profit-Agnostic Targeting Models

Targeting models belong to the field of supervised learning (Hastie et al., 2009). Assume a marketer wishes to predict the behavior of customer i, characterized by $\mathbf{x}_i = (x_{1i},\ x_{2i}, \ldots, x_{\mathrm{Mi}}) \in \mathbb{R}^M$, where the elements of $\boldsymbol{x}_i$ capture transactional and demographic information. Let y_{i} denote the response of customer i to a past marketing action. The response may be continuous (e.g., purchase amount) or discrete (e.g., whether an offer was accepted). We focus on binary classification where $y_i \in \{0, 1\}$, with a value of y_{i}=1 (y_{i}=0) indicating that customer i accepted (rejected) a marketing offer. A targeting model, $f(\boldsymbol{x})$, represents a functional mapping from customer records to responses: $f_{\boldsymbol{\Lambda}}(\mathbf{x}) : \mathbb{R}^M \longmapsto \{0, 1\}$, where Λ denotes a vector of model parameters. Model estimation involves fitting model parameters to data. Afterwards, the model allows the marketer to predict customer response (and more generally behavior) from observable customer data.

Targeting model development follows an inductive approach: Given a data set of customer records and corresponding responses, $D = (y_i, \mathbf{x}_i)_{i=1}^N$, a learning algorithm fits the model parameters, Λ, so as to minimize the deviation between model estimates and actual responses: $\boldsymbol{\Lambda}' \leftarrow Q\left(y_i, f_{\boldsymbol{\Lambda}}(\mathbf{x}_i)\right) \quad \forall\ i = 1,\ \ldots,\ N$, where $\boldsymbol{\Lambda}'$ denotes the optimal set of parameters and the loss function Q measures the disagreement between model outputs and data. Therefore, model estimation is equivalent to minimizing a loss function over D. A loss function represents a model-internal notion of fit. Considering the logit model as an example, Q equals the negative log-likelihood (NLL). Common statistical loss functions (NLL, cross-entropy, Hinge loss, etc.) implement the principles of statistical learning to ensure that a model is able to generalize to novel data. Prediction models estimated using such loss functions are generic and can be employed in many domains. However, they disregard specific application characteristics unless these are accurately reflected in the loss function. We argue that a close correspondence between a model-internal internal notion of fit and business value should not be taken for granted. Maximizing fit using some statistical loss function may lead to a different model compared to maximizing campaign profit. On the other hand, statistical loss functions have strong theoretical underpinnings and exhibit desirable properties related to generalization (Hastie et al., 2009). It is imperative to build on this theory when developing a prediction model. This motivates our PCES approach to integrate statistical considerations (in the form of established loss functions and estimation principles) and business value (in the form of campaign profit) during target model development.

3.3.2 Target Group Selection and Model Assessment in Marketing Campaign Planning

Campaign planning aims at maximizing the efficiency of resource utilization. Contacting customers with a marketing message entails a cost so that it is typically inefficient to target the whole customer base. Instead, marketers use targeting models to estimate response probabilities on a customer level. This facilitates restricting solicitations to likely responders. Applications of targeting models include the mail-order industry, churn management, and cross-selling. Recently, targeting models are increasingly used in real-time settings such as digital marketing (Perlich et al., 2014) and social media (Li et al., 2018).

From a managerial point of view, the business value of a targeting model depends on the degree to which it increases the profitability of targeted marketing actions. We model the profit of a marketing campaign, Ω, as follows (Martens & Provost, 2011):

$$\Omega(l(\tau), \tau) = N \cdot \tau \cdot (\pi_+ \cdot l(\tau) \cdot r - c), \tag{3.1}$$

where N denotes the size of the customer base, τ the fraction of targeted customers (i.e., campaign size), and π_+ the base rate of customers willing to accept the marketing offer in the customer base. The parameters r and c represent the return and cost associated with an accepted offer and making the offer, respectively. The quantity $l(\tau)$, called the lift, is a marketing specific measure of predictive accuracy, which depends on the size of the campaign, τ. With π_τ denoting the fraction of responses in the target group, the lift is given as:

$$l(\tau) = \frac{\pi_\tau}{\pi_+} \tag{3.2}$$

A campaign that targets customers at random reaches a fraction of π_+ actual responders. Thus, the lift assesses the degree to which a model-based targeting improves over a random benchmark.

Revised versions of Eq. 3.1 have been proposed to capture the characteristics of specific marketing applications. Neslin et al. (2006) devise a profit function for models that target retention actions to customers with high churn probability. The expected maximum profit criterion further refines this approach (Verbraken et al., 2012). The advantage of the campaign profit function Eq. 3.1 over subsequent advancements is generality. Connecting customer revenues, direct costs, and model accuracy through model lift, Eq. 3.1 can represent a variety of targeting applications including churn management, direct mail, e-couponing, etc. Therefore, we use Eq. 3.1 in this paper and leave the evaluation of the proposed PCES approach for specific targeting tasks such as churn modeling to future work.

An assumption of Eq. 3.1 and its extensions is that costs and returns are homogeneous across customers. In campaign planning, assuming constant offer costs is plausible for most marketing channels. However, disregarding variability in customer spending (r=const.) is a strong

simplification. Typically, the returns from accepted marketing offers differ across customers. Our justification for using Eq. 3.1 despite this assumption is threefold. First, it is common practice to work with class as opposed to case depending costs/returns in the marketing and cost-sensitive learning literature (Rokach et al., 2008; Verbeke et al., 2012). Second, calculating campaign profit using the mean revenue per accepted offer may be more suitable for predictive modeling, for example because information to estimate revenues at the customer level reliably is lacking. Last, some applications do not require distinguishing revenues across customers, for example when targeting services entail a fixed fee or when running lead generation campaigns.

3.3.3 Profit-Conscious Ensemble Selection

The proposed modeling framework is based on the view that the development of predictive decision support models should pay attention to both statistical and business considerations. Therefore, we strive to incorporate campaign profit Eq. 3.1 as marketing objective into model development (see Figure 3.1). To achieve this, we decompose model development into two sub-steps. The first stage leverages statistical learning principles. In step two, model predictions are refined to maximize campaign profit. Recall that such multi-stage approach mimics the way in which managers use decision support models: they re-appraise and possibly correct DSS outputs in the context of their decision task (Fuller & Dennis, 2009).

The proposed framework is based on a machine learning paradigm called ensemble selection (Caruana et al., 2006). An ensemble is a collection of (base) models, all of which predict the same target. Much research confirm that combining multiple models in an ensemble is useful to increase predictive accuracy (Verbeke et al., 2012). Ensemble selection involves three steps: i) constructing a library of candidate models (*model library*), ii) selecting an "appropriate" subset of models for the ensemble (*candidate selection*), and iii) integrating the predictions of the chosen models into a composite forecast (*model aggregation*). From an algorithmic point of view, PCES follows Caruana's et al. (2006) approach. Its distinctive feature is that it integrates statistical and economic objectives. This way, PCES embodies a different paradigm toward developing predictive decision support models.

Model Library

The success of an ensemble depends on the diversity of its members. To obtain a library of diverse models, we use different learning algorithms. We also consider multiple settings for algorithmic meta-parameters. Meta-parameters such as the regularization parameter in support vector machines facilitate adapting a learning algorithm to a task, which suggests that prediction models from the same algorithm vary with meta-parameters and display diversity. Table 3.1 summarizes the learning algorithms and meta-parameter settings in the model library.

It is common practice to select a specific, 'best' set of meta-parameters for an individual learning algorithm in a model selection step. As we detail below (see Section 4.2), we also adopt this practice to obtain benchmark models against which we compare PCES. However, for PCES itself, we do not perform model selection a priori but keep all candidate models in the library.

The selection of algorithms and meta-parameters is based upon previous literature on customer targeting and ensemble modeling (Lessmann et al., 2015; Verbeke et al., 2012). Some methods have been chosen due to their popularity (e.g., logistic regression, decision trees, discriminant analysis) and others because of high performance in previous studies (e.g., random forest, support vector machines, gradient boosting). Interested readers can find a comprehensive discussion of the algorithms in Hastie et al. (2009). In total, we consider 15 learning algorithms from which we derive 877 different models. We acknowledge that several extensions of popular machine learning algorithms have been proposed in the literature. Innovative learners like, for example, the fuzzy support vector machine (Wang et al., 2017) may give better results than the original version of the algorithm. Our reason to not include corresponding techniques comes from the design goal of PCES to be easy to implement in practice. Standard algorithms as those forming our model library are available in contemporary business analytics software such as, e.g., SAS, Microsoft Azure ML, and many others as well as popular data science programming languages such as R, Python, Scala, etc. or high-performance computing infrastructures like Apache Spark. Leveraging corresponding standards is beneficial because it ensures that companies could deploy PCES at low cost and without a need to re-implement algorithms that have mainly been used in research. The same consideration discourages an application of deep learning in this paper.

Table 3.1: Classification methods and meta-parameter settings

Learning Algorithm	Meta-parameter*	Candidate Settings**
Classification and Regression Tree Recursively partitions a training data set by inducing binary splitting rules to minimize the impurity of child nodes in terms of the Gini coefficient. Terminal nodes are assigned a posterior class-membership probability according to the distribution of the classes of the training instances in this node. To classify novel instances, the splitting rules learned during model building are employed to determine an appropriate terminal node.	Min. size of nonterminal nodes Pruning of fully grown tree *Overall number of models: 6*	10, 100, 1000 Yes, No
Artificial Neural Network Three-layered architecture of information processing-units referred to as neurons. Each neuron receives an input signal in the form of a weighted sum over the outputs of the preceding layer's neurons. This input is transformed by a logistic function and passed to the next layer. The neurons of the first layer are the covariates of a classification task. The output layer consists of a single neuron, whose output can be interpreted as a class-membership probability. Building a neural network model involves determining connection weights by minimizing a regularized loss-function over training data.	No. of neurons in hidden layer Regularization factor (weight decay) *Overall number of models: 162*	3, 4, . . . , 20 $10^{[-4,-3.5,\ldots,0]}$
Naive Bayes Approximates class-specific probabilities under the assumption that all covariates are statistically independent.	Histogram bin size *Overall number of models: 9*	2, 3, . . . , 10
k-Nearest-Neighbor Decision objects are assigned a class-membership probability according to the class distribution prevailing among its k nearest (in terms of Euclidian distance) neighbors.	Number of nearest neighbors *Overall number of models: 18*	10, 100, 150, 200, . . . , 500, 1000, 1500, . . . , 4000

Table 3.1: Classification Methods and Meta-Parameter Settings (cont.)

Learning Algorithm	Meta-parameter*	Candidate Settings**
Linear Discriminant Analysis Approximates class-specific probabilities by means of multivariate normal distributions assuming identical covariance matrices. This assumption yields a linear classification model, whose parameters are estimated by means of maximum likelihood procedures.	Covariates considered in the model *Overall number of models: 20*	Full model, stepwise variable selection with p-values in the range 0.05, 0.1,. . . , 0.95
Logistic Regression Approximates class membership probabilities (i.e., a posteriori probabilities) by means of a logistic function, whose parameters are estimated from training data by maximum likelihood procedures.	Covariates considered in the model *Overall number of models: 20*	Full model, stepwise variable selection with p-values in the range 0.05, 0.1,. . . , 0.95
Quadratic Discriminant Analysis Differs from LDA only in terms of the assumption about the structure of the covariance matrix. Relaxing the assumption of identical covariance leads to a quadratic discriminant function.	Covariates considered in the model *Overall number of models: 20*	Full model, stepwise variable selection with p-values in the range 0.05, 0.1,. . . , 0.95
Regularized Logistic Regression Differs from ordinary LogR in the objective function optimized during model building. A complexity penalty given by the L1-norm of model parameters (Lasso-penalty) is introduced to obtain a "simpler" model.	Regularization factor *Overall number of models: 29*	$2^{[-14,-13,...,14]}$
Support Vector Machine with Linear Kernel Constructs a linear boundary between training instances of adjacent classes so as to maximize the distance between the closest examples of opposite classes and achieve a pure separation of the two groups.	Regularization factor *Overall number of models: 29*	$2^{[-14,-13,...,14]}$

Table 3.1: Classification Methods and Meta-Parameter Settings (cont.)

Learning Algorithm	Meta-parameter*	Candidate Settings**
Support Vector Machine with Radial Basis Function Kernel Extends linear SVM by implicitly projecting training instances to a higher dimensional space by means of a kernel function. The linear decision boundary is constructed in this transformed space resulting in a nonlinear classification model.	Regularization factor Width of Rbf kernel function *Overall number of models: 300*	$2^{[-12,-13,...,12]}$ $2^{[-12,-11,...,-1]}$
AdaBoost Constructs an ensemble of decision trees in an incremental manner. The new members to be appended to the collection are built in a way to avoid the classification errors of the current ensemble. The ensemble prediction is computed as a weighted sum over the member classifiers' predictions, whereby member weights follow directly from the iterative ensemble building mechanism.	No. of member classifiers *Overall number of models: 11*	10, 20, 30, 40, 50, 100, 250, 500, 1000, 1500, 2000
Bagged Decision Trees Constructs multiple CART trees on bootstrap samples of the original training data. The predictions of individual members are aggregated by means of average aggregation.	No. of member classifiers *Overall number of models: 11*	10, 20, 30, 40, 50, 100, 250, 500, 1000, 1500, 2000
Bagged Neural Networks Equivalent to Bagged DT but using ANN instead of CART to construct member classifiers. The ensemble prediction is computed as a simple average over member predictions.	No. of member classifiers *Overall number of models: 5*	5, 10, 25, 50, 100
LogitBoost Modification of the AdaB algorithm which considers a logistic loss function during the incremental member construction. We employ tree-based models as member classifiers.	No. of member classifiers *Overall number of models: 11*	10, 20, 30, 40, 50, 100, 250, 500, 1000, 1500, 2000

Table 3.1: Classification Methods and Meta-Parameter Settings (cont.)

Learning Algorithm	Meta-parameter*	Candidate Settings**
Random Forest An ensemble of fully grown CART derived from bootstrap samples of the training data. In contrast with standard CART that determine splitting rules over all covariates, a subset of covariates is randomly drawn whenever a node is branched and the optimal split is determined for these preselected variables. The additional randomization increases diversity among member classifiers. The ensemble prediction follows from average aggregation.	No. of member classifiers	100, 250, 500, 750, 1000, 1500, 2000 ***
	No. of covariates randomly selected for node splitting	$[0.1, 0.5, 1, 2, 4] \cdot \sqrt{M}$***
	Overall number of models: 35	
Stochastic Gradient Boosting Modification of the AdaB algorithm, which incorporates bootstrap sampling and organizes the incremental ensemble construction in a way to optimize the gradient of some differential loss function with respect to the present ensemble composition. We employ tree-based models as member classifiers.	No. of member classifiers	10, 20, 30, 40, 50, 100, 250, 500, 1000, 1500, 2000
	Overall number of models: 11	

* Note that Table 3.1 depicts only those meta-parameters for which we consider multiple settings. A classification method may offer additional meta-parameters.

** We consider all possible combination of meta-parameter settings for learners such as Random Forest that exhibit multiple meta-parameters.

*** M represents the number of explanatory variables (i.e., covariates) in a data set.

Candidate Selection

Given the model library, we select candidate models using directed hill-climbing (Caruana et al., 2006). In particular, we first select the single best candidate model from the library. To improve this model's performance, we next assess all pairwise combinations of the chosen model and one other base model from the library. This way, we obtain a collection of possible two-member ensembles, out of which we select the best performing candidate ensemble. We then continue with examining the set of all three-member ensembles that include the models chosen in the previous iteration. Incremental ensemble growing terminates when adding novel members stops improving performance. Interested readers find a working example of the algorithm in the e-companion (see Appendix 3.A).

We propose to reserve the selection step for business objectives. Using heuristic search, it is possible to gear ensemble selection toward any objective function that depends on the model-estimated probabilities. In this paper, we devise an ensemble that incorporates business objectives through maximizing Eq. 3.1 in the selection stage. This way, PCES refines the first-stage predictions, which stem from well-established prediction models and embody the principles of statistical learning, by means of a combination of predictions to better represent the actual decision problem.

From a mathematical point of view, configuring the hill-climbing heuristic to maximize Eq. 3.1 appears a minor modification. However, this modest modification leads to a fundamentally different paradigm toward prediction model development. The campaign profit function Eq. 3.1 captures the business value of a decision support model and characteristics of the decision context such as a budget constraint (in the form of τ. Consequently, maximizing Eq. 3.1 leads to a contextualized model that is aware of the environment to which it will be deployed and the decisions it is meant to support. Furthermore, an ex-post revision of (individual model) predictions as done by PCES mimics the way in which managers use DSS recommendations and possibly correct for misleading advice (Fuller & Dennis, 2009). These features represent the real value of PCES and, as we suggest, warrant a comprehensive empirical evaluation how much a contextualized modeling paradigm improves over standard supervised learning.

Model Aggregation

Model aggregation refers to a combination of models' predictions. PCES combines predictions in the candidate selection step. A candidate ensemble consists of a subset of base models. To assess a candidate ensemble, we compute the simple average over the predictions of the selected base models. We detail this approach in the appendix, which provides a numerical example of candidate selection and PCES (see Appendix 3.A). PCES performs the same model aggregation when computing the predictions of the final ensemble, which is the specific selection of base models that gives the best results during candidate selection.

Although we pool models by averaging over their predictions, PCES effectively computes a weighted average. This is because the candidate selection procedure of Caruana et al. (2006)

allows the same model to enter the ensemble multiple times. The opportunity to weight predictions whenever the data suggest that a strong model deserves greater influence on the ensemble prediction adds to the flexibility of ensemble selection. Note that averaging model predictions requires all models to produce forecasts of a common scale. To ensure this, we calibrate base model predictions using a logistic link function prior to model averaging (Platt, 2000).

3.4 Empirical Design

We examine the effectiveness of PCES in the scope of an empirical benchmark. Such experiment requires suitable data, which represents the characteristics of customer targeting applications, and benchmark models to put the performance of PCES into context.

3.4.1 Marketing Data Sets

The empirical study considers 25 cross-sectional marketing data sets. The data sets stem from different industries and represent different prediction tasks, each of which requires selecting customers for targeted marketing actions. The main sources from which we gather the data sets are: i) data mining competitions, ii) previous modeling studies, iii) the UCI machine learning repository (Lichman, 2013), and iv) projects with industry partners. Given the large number of data sets, it is prohibitive to discuss every data set in detail. Table 3.2 summarizes data set characteristics and identifies sources where more information is available. Every data set has been recorded at a given point in time. Accordingly, variable values give a snapshot of the state of a customer but provide no information how a variable, say customer spending, has evolved over time. For this reason, we do not consider sequence learning algorithms such as recurrent neural networks in this paper.

To simulate a real-world campaign planning setting, we randomly split data sets into two samples using a ratio of 60:40. We refer to the two samples as the training set and the test set, respectively. We develop targeting models using the training set and assess fully specified models on the test set. Certain modeling choices within PCES and the benchmark models (see below) require auxiliary validation data. Examples include the identification of the best base model in the library (as benchmark to PCES) and the heuristic search for ensemble members in the second stage of PCES. We obtain such validation data by means of five-fold cross validation on the training set (Caruana et al., 2006).

3.4.2 Benchmark Models

Alternative targeting models represent a natural benchmark to the proposed PCES approach. We consider i) the well-known logit model, due to its popularity in marketing , ii) random forest, due to its success in previous benchmarking studies (Lessmann et al., 2015; Verbeke et al., 2012), and iii) a best base model (BBM) benchmark, which is given by the strongest individual targeting model from the model library. A common denominator among these benchmarks is that they account for the problem context during *model selection*. For each marketing data set, we select among the 20 / 35 / 877 candidate logit / random forest / base models (see Table

Table 3.2: Data set characteristics

Data	Marketing Goal	Industry	Source*	Obs.	Var.	P(+1)**
D1	Churn prediction	Energy	DMC02	20,000	32	0.10
D2	Churn prediction	Finance	CP	155,056	23	0.14
D3	Churn prediction	Finance	CP	30,104	47	0.04
D4	Churn prediction	Telco	[37]	40,000	70	0.50
D5	Churn prediction	Telco	[37]	93,893	196	0.50
D6	Churn prediction	Telco	[37]	12,410	18	0.39
D7	Churn prediction	Telco	[37]	69,309	67	0.29
D8	Churn prediction	Telco	[37]	21,143	384	0.12
D9	Churn prediction	Telco	KDD09	50,000	301	0.07
D10	Churn prediction	Telco	[37]	47,761	41	0.04
D11	Churn prediction	Telco	[37]	5,000	18	0.14
D12	Profitability scoring	E-Commerce	DMC05	50,000	119	0.06
D13	Profitability scoring	E-Commerce	DMC06	16,000	24	0.49
D14	Profitability scoring	Mail-order	UCI-Adult	48,842	17	0.24
D15	Profitability scoring	Mail-order	DMC04	40,292	107	0.21
D16	Response modeling	Charity	KDD98	191,779	43	0.05
D17	Response modeling	E-Commerce	CP	121,511	82	0.06
D18	Response modeling	E-Commerce	CP	214,709	77	0.13
D19	Response modeling	E-Commerce	CP	382,697	76	0.09
D20	Response modeling	E-Commerce	DMC10	32,428	40	0.19
D21	Response modeling	Finance	CP	45,211	16	0.12
D22	Response modeling	Finance	UCI-Coil	9,822	13	0.06
D23	Response modeling	Mail-order	DMC01	28,128	106	0.50
D24	Response modeling	Publishing	CP	300,000	30	0.01
D25	Response modeling	Retail	DMC07	100,000	17	0.24

*CP = consultancy project with industry; DMC[year] = Data Mining Cup (http://www.data-mining-cup.com); KDD[year] = ACM KDD Cup (http://www.sigkdd.org/kddcup/index.php); UCI-xxx = UCI Machine Learning Repository[2] (with xxx being the name of the data set in the repository).

**P(+1) denotes prior probability of response (e.g., fraction of customers who accept an offer).

3.1) the one giving maximal campaign profit Eq. 3.1. Prior work finds a selection of prediction models using business performance measures to substantially improve decision quality (Glady et al., 2009; Verbeke et al., 2012; Verbraken et al., 2014). Therefore, we expect the benchmarks to be challenging. To further elaborate on our approach toward benchmark selection, recall that our model library includes multiple models for each learning algorithm, which we derive from executing the algorithm with different settings for algorithmic meta-parameters (see Table 3.1). We select the logit and random forest benchmarks among all logit and random forest models in the model library for each data set and for each experimental setting. For example, we consider multiple cost-to-benefit ratios and examine model performance across these ratios on each data set. We also consider different mailing depths. In the interest of obtaining a challenging benchmark, we select the strongest logit/random forest model for each setting and data set individually. We proceed in the same way to select the BBM, this time, however not selecting the benchmark model only among candidate logit / random forest models but all models in the library.

The ensemble selection approach of Caruana et al. (2006) contributes a fourth benchmark. Here, we call it profit-agnostic ensemble selection (PAES) and employ a statistical loss function (i.e., NNL) for base model selection. Therefore, PAES and PCES differ in their approach to select base models the the final ensemble in a profit-agnostic as opposed to a profit-conscious manner. This configuration allows us to attribute performance differences between PAES and PCES to the fact that the latter accounts for business performance during model development.

The last benchmark draws inspiration from Bhattacharyya (1999). It optimizes the coefficients of a linear regression function, which discriminates between responsive and non-responsive customers, using a genetic algorithm (GA). We use Eq. 3.1 as fitness function implying that the GA maximizes campaign profit. Focusing exclusively on business goals during model development, GA is a useful benchmark to support the design of PCES as an integrated modeling framework that balances statistical and economic considerations. GAs exhibit meta-parameters such as the size of the population, the specific type of crossover operator or the mutation rate. In configuring the GA benchmark, we rely on Bhattacharyya (1999) and use their settings of population size=50, crossover rate=0.7, and mutation rate=0.2.

3.4.3 Configuration of Ensemble Selection

Caruana et al. (2006) propose some modifications of basic ensemble selection. One extension consists of an additional bagging step. Instead of selecting a single set of base models from the full model library, they subsample the library, select one ensemble from each subsample, and average over the resulting ensembles (Caruana et al., 2006). The basic and bagged ensemble selection algorithms represent alternative strategies to develop a model. We consider both strategies and determine the superior approach for each data set by means of model selection. For bagged ensemble selection, we consider subsample sizes of 5, 10, and 20 percent of the model library and 5, 10, and 25 bagging iterations. Importantly, PAES and PCES are treated in the same way to avoid bias.

3.5 Empirical Results

The experimental design provides test set predictions from PCES and benchmark models across the marketing data sets. Many indicators are available to assess predictive accuracy. We suggest that a comparison in terms of business performance is most meaningful from a managerial point of view and thus assess targeting models in terms of campaign profit Eq. 3.1.

Recall that Eq. 3.1 is a function of campaign size, τ. In the following, we consider τ a decision variable and let a targeting model find the profit maximal solution to Eq. 3.1 over $l(\tau)$ and τ. This implies that the model determines which and how many customers to target and thus how much to spend on the campaign. Verbeke et al. (2012) recommend this approach and proof its effectiveness. We follow their advice but consider a different profit function to cover a larger scope of marketing applications.

To cover a broad range of application scenarios, we consider multiple settings for the monetary campaign parameters offer cost (c) and return per accepted offer (r). More specifically, it is sufficient to vary r because the profit function Eq. 3.1 is invariant to a linear scaling. Rescaling Eq. 3.1 such that c=1 and r'=r/c does not change the profit maximal solution. We thus fix c at \$1 and consider settings of r = \$2, \$3, \$5, \$10, \$15, \$25, \$50, \$75, and \$100. These values capture a range of targeting applications. Smaller values represent settings where the ratio between offer cost and return per accept is moderately skewed. Such scenario might occur when companies contact customers through a call-center or when selling products by means of printed catalogs in the mail-order industry. Both channels involve considerable offer costs (e.g., to produce a premium catalog), which could explain moderate imbalance between r and c. High skewness between these parameters arises in online marketing where digital channels facilitate reaching customers at very low costs. Larger values of r capture such applications. Overall, considering 25 marketing data sets with 11 settings for the cost-to-benefit ratio, r/c, we obtain 275 experimental settings. To carry out profit optimization, we run PCES as well as the PAES and GA benchmark for each of these settings individually. For the logit, random forest and BBM benchmark, we use the models stored in the model library and respectively select the best logit, random forest, and base model for each experimental setting. Given that larger values of r give an incentive to increase campaign size, we constrain the optimization of Eq. 3.1 such that $\tau \leq 0.5$. Since marketing campaigns typically target a small fraction of customers, contacting more than half of the customer base seems unrealistic.

Table 3.3 reports the win-tie-loss statistics of PCES vs. benchmark models for the 11 (return to cost ratios) $\times$ 25 (data sets) = 275 comparisons. Consider, for example, the comparison of PCES versus BBM at r=\$2. A value of 22 suggests that PCES achieves higher campaign profit than BBM on 22 out of 25 data sets. BBM outperforms PCES on two data sets and both models tie on one data set. We also compare the statistical significance of profit differences using the Friedman test (see bottom of Table 3.3). For the results of Table 3.3, a χ^2 value of 823.5 indicates that we can reject the null hypothesis of equal performance (p-value <0.000). This allows us to proceed with a set of pairwise comparisons of PCES against one benchmark to detect significant differences among individual targeting models. To protect against an

Table 3.3: Win-tie-loss statistics of PCES versus benchmarks in the flexible budget case

Return	PCES vs. Logit			PCES vs. RF			PCES vs. BBM			PCES vs. GA			PCES vs. PAES		
	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss
\$2	24	1	0	21	2	2	22	1	2	25	0	0	19	3	3
\$3	24	0	1	21	1	3	22	1	2	25	0	0	22	0	3
\$4	25	0	0	24	0	1	21	1	3	25	0	0	20	0	5
\$5	25	0	0	23	1	1	23	1	1	24	1	0	20	0	5
\$10	24	0	1	24	0	1	22	0	3	24	0	1	18	0	7
\$15	24	0	1	23	0	2	18	0	7	24	0	1	12	0	13
\$20	24	0	1	23	0	2	22	0	3	24	0	1	17	0	8
\$25	24	0	1	24	0	1	23	0	2	23	0	2	16	1	8
\$50	23	0	2	23	0	2	22	0	3	24	0	1	16	0	9
\$75	23	0	2	21	1	3	21	0	4	24	0	1	13	0	12
\$100	23	0	2	19	1	5	20	0	5	23	1	1	11	1	13
Total	263	1	11	246	6	23	236	4	35	265	2	8	184	5	86
	96%	0%	4%	89%	2%	8%	86%	1%	13%	96%	1%	3%	67%	2%	31%
p-value*	0.000			0.000			0.000			0.000			0.000		

* The p-values correspond to pairwise comparisons of PCES and one benchmark, using Rom's procedure to protect against an elevation of the α error in multiple pairwise comparisons (García et al., 2010). Multiple pairwise comparisons are feasible since a χ^2 value of 823.5 suggest that we can reject the null hypothesis of equal performance among models (Friedman test) with high confidence (p-value < 0.000).

elevation of alpha values in multiple pairwise comparisons, we adjust p-values using Rom's procedure (García et al., 2010). The last row of Table 3.3 reports the adjusted p-values.

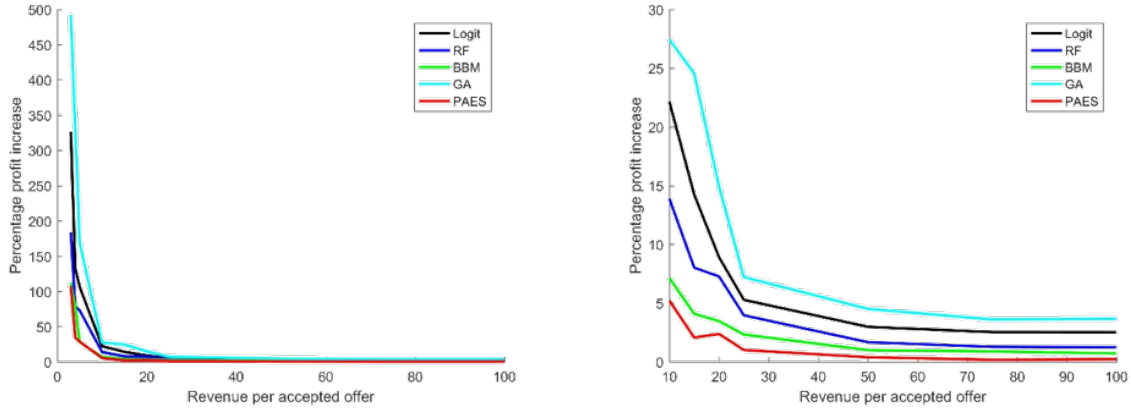
Table 3.3 reveals evidence that PCES produces significantly higher campaign profits than any of the benchmark models (p-values of pairwise comparisons consistently less than 0.000). Recall that the purpose of the logit, RF, and BBM benchmark is to reflect common marketing practices where a set of candidate models is developed and the strongest candidate (in terms of Eq. 3.1) is selected. This is exactly the modeling paradigm advocated in previous studies (Verbeke et al., 2012; Verbraken et al., 2012). Accordingly, the results of Table 3.3 indicate that introducing the relevant notion of model performance during model development (as opposed to model selection) further increases performance. However, this interpretation requires further qualification since the superiority of PCES may also come from the ability of ensemble selection to create powerful prediction models. Indeed, the PAES benchmark, an ordinary ensemble selection method, turns out to be the strongest benchmark. However, although benefitting from the same large base model library as PCES, a PAES-based customer targeting gives significantly less profit compared to using PCES. In particular, we find the latter to produces higher profits in 184 out of 275 comparisons (67 percent). Before examining the relative performance of alternative targeting models in more detail, we note that PCES also outperforms the GA benchmark (i.e., a direct profit maximization) with substantial margin.

To obtain a clearer view on the degree to which PCES increases business performance, we calculate the profit implication resulting from using PCES or a benchmark model for campaign

Table 3.4: Comparison of campaign profit at model-optimized campaign sizes

	Campaign profit [$]					
Data	**Logit**	**RF**	**BBM**	**GA**	**PAES**	**PCES**
D1	1,660	1,596	1,764	1,532	**1,874**	1,846
D2	61,612	75,816	75,989	62,953	75,725	**76,001**
D3	-2	-83	88	-104	76	**137**
D4	-2,992	-2,832	-2,832	-3,052	-2,852	**26**
D5	-7,096	-6,766	-6,766	-7,096	-6,666	**25**
D6	-1,017	-997	-977	-1,027	-997	**159**
D7	35,578	39,598	39,778	35,098	40,408	**40,618**
D8	2,966	2,926	3,270	2,756	**3,404**	3,121
D9	699	469	862	509	999	**1,139**
D10	442	876	839	590	901	**984**
D11	1,491	2,000	2,022	1,534	2,020	**2,058**
D12	-8	17	-33	-310	84	**428**
D13	14,700	18,270	18,270	15,110	18,390	**18,810**
D14	34,421	34,755	35,067	34,385	35,107	**35,185**
D15	21,642	21,842	**22,012**	21,353	21,982	21,073
D16	572	6	572	208	527	**726**
D17	9,121	9,283	9,690	9,568	**10,690**	10,087
D18	64,096	101,186	105,824	63,438	105,649	**106,418**
D19	85,123	119,158	122,949	91,387	**123,881**	123,804
D20	10,424	10,614	10,564	9,954	10,654	**10,884**
D21	12,877	14,534	14,632	12,708	14,498	**14,725**
D22	210	323	325	242	305	**357**
D23	29,044	29,544	**30,154**	28,454	30,074	30,004
D24	-1	-2	14	1	13	**27**
D25	47,440	53,210	53,210	50,380	**53,770**	53,660
Estimated profit increase (in percent)*	657 (22%)	407 (14%)	233 (7%)	756 (27%)	178 (5%)	

*The estimation is based on García et al. (2010). We first use their contrast estimation approach to calculate the expected profit improvement of PCES over a benchmark, and then convert this contrast to a percentage through dividing by the benchmark's median (across data sets) campaign profit.



(a) All settings of return r

(b) Filtered by return $r>5$ for readability

Figure 3.2: Expected percentage improvement in campaign profit due to using PCES for target group selection. We estimate profit contrasts in the same way as in Table 3.4

Table 3.5: Model-optimized campaign sizes

	Model-optimized campaign sizes [%]					
Data	**Logit**	**RF**	**BBM**	**GA**	**PAES**	**PCES**
D1	41.12	49.68	35.58	40.09	38.20	43.18
D2	25.78	16.21	15.67	26.15	15.49	15.86
D3	0.35	6.67	4.33	4.01	7.25	4.76
D4	50.00	50.00	50.00	50.00	50.00	0.17
D5	50.00	50.00	50.00	50.00	50.00	0.34
D6	50.00	50.00	50.00	50.00	50.00	1.97
D7	50.00	50.00	50.00	50.00	50.00	50.00
D8	46.16	47.70	46.34	46.87	49.26	50.00
D9	7.70	12.70	16.04	13.20	23.10	16.96
D10	5.07	6.56	5.81	5.44	7.69	5.74
D11	38.43	15.47	14.40	39.77	14.00	15.10
D12	14.14	15.26	17.36	12.35	16.18	7.86
D13	50.00	50.00	50.00	50.00	50.00	50.00
D14	48.52	49.62	48.59	49.83	48.85	47.68
D15	50.00	50.00	50.00	49.93	50.00	45.34
D16	3.83	0.03	3.83	0.71	2.57	4.27
D17	22.04	17.39	17.61	15.44	19.52	16.66
D18	36.83	20.09	17.74	35.45	17.56	17.03
D19	19.52	13.03	12.14	18.99	12.55	12.04
D20	50.00	50.00	50.00	50.00	50.00	50.00
D21	28.99	25.47	26.97	30.64	25.78	27.95
D22	23.65	15.44	14.63	18.51	23.02	10.77
D23	50.00	50.00	50.00	50.00	50.00	50.00
D24	0.00	0.01	0.04	0.06	0.04	0.04
D25	50.00	50.00	50.00	50.00	50.00	50.00
Median	38.43	25.47	26.97	39.77	25.78	16.66

targeting. In particular, we consider a fictitious company with a customer base of $N = 100{,}000$ customers; and let the per-customer return from accepted offers, r, and offer costs to contact customers, c, be \$10 and \$1, respectively. Table 43.4 depicts the campaign profits emerging from a model-based targeting per marketing data set. Given that we consider campaign size a decision variable, we let every targeting model select its individually best setting τ. This way, Table 3.4 compares targeting models in terms of the maximal campaign profit they can produce for given r and c. Bold face highlights the best result per data set. The optimized campaign sizes corresponding to the results of Table 3.4 are available in Table 3.5. The last row of Table 3.4 summarizes the observed results in the form of an estimate of the expected profit increase of PCES over a benchmark. The estimation procedure comes from García et al. (2010) and is based on the median profit difference between PCES and a benchmark model across the data sets. Given the scope of the empirical study (e.g., 25 real-world data sets from different industries), we consider the resulting value a reliable estimate of the profit that a targeting model achieves on unseen data.

Table 3.4 reemphasizes that PCES typically produces higher profits than benchmark models. This is especially apparent when examining the performance contrast shown in the last row of Table 3.4. Based on the observed results, we expect PCES to increase campaign profit by five percent compared to the most challenging benchmark and up to fourteen percent compared to random forest, a state-of-the-art classifier much credited for high accuracy (Lessmann et al., 2015). Profit increases of five percent and above are managerially meaningful, especially for larger companies and companies that run many campaigns (Neslin et al., 2006). It is also noteworthy that using the logit model for targeting, an approach still popular in industry, entails substantial opportunity costs. Compared to this benchmark, PCES produces higher campaign profits across all data sets and can be expected to increase profits by 22 percent on average. With respect to a direct optimization of campaign profit during model development, which the GA benchmark embodies, Table 3.4 reveals that corresponding results are the weakest in the comparison. Last, PCES is the only approach that avoids losses. For some data sets (e.g., D4-D6) the optimization of τ on validation data gives a poor result for the hold-out test data on which we calculate campaign profit. In particular, Table 3.5 reveals that all benchmarks select τ equal to its upper bound of 0.5 on D4 - D6. This leads to large campaigns that result in a loss for the given setting of $r{:}c = 10{:}1$. PCES, on the other hand, benefits from its ability to adapt the ensemble forecast when optimizing τ, because it employs Eq. 3.1 during model development. This allows PCES to recognize that the level of predictive accuracy vis-a-vis the return to cost ratio might not facilitate profitable targeting. Thus, PCES selectsτ close to zero. Finally, Table 3.5 evidences a trend of PCES to recommend smaller campaigns. The median value τ =16.66 for PCES is much less than the second-smallest value of τ =25.47 for RF. Smaller campaigns are appealing since they require less resources and might be better targeted to customer interests. For example, despite recommending smaller campaigns, PCES produces higher profits than RF on all data sets, which signals higher predictive accuracy and, in turn, better targeting.

The results of Table 3.4 and Table 3.5 stem from a campaign with specific setting of returns

and offer costs. To confirm generalizability of results to other campaign settings, we next examine the magnitude of PCES-induced profit improvements across the full range of campaign parameters r = \$2, \$3, \$5, \$10, \$15, \$25, \$50, \$75, and \$100 (with c = \$1). To that end, we rerun model development (for PCES and GA) and model selection (logit, RF, BBM, PAES) for all data sets and settings of r. We then use the same contrast estimation approach (see last row Table 3.4) to calculate percentage profit improvements of PCES over its benchmarks (García et al., 2010). Figure 3.2 depicts the corresponding results. Given that smaller settings of r lead to large improvements over weaker benchmarks, we split Figure 3.2 into two panels which show results for all settings of r and those above five, respectively.

Figure 3.2 confirms that superior performance of PCES generalizes to other settings of campaign parameters. Above zero improvements demonstrate that PCES consistently produces higher profits than the benchmarks. GA is again the weakest benchmark in the comparison. Even in the scenario r:c=100:1, where high imbalance between marketing returns and costs renders the targeting task relatively easy, PCES increases campaign profits by more than five percent compared to GA. This confirms that direct maximization of campaign profits is not a suitable approach to develop targeting models. The other models ground on statistical learning. From Figure 3.2, we conclude that following corresponding principles is essential when developing a targeting model. However, the specific adaptation that we propose, namely to introduce campaign profits into model development, succeeds in improving the business performance of the resulting model. Random forest, for example, recommends campaigns that are roughly 3–15% less profitable compared to PCES.

3.6 Discussion

The empirical analysis evidences the effectiveness of the proposed approach toward model development. Our study also sheds light on the divergence between the optimization of statistical loss and business objectives for prediction model development in targeting applications. The experimental design includes three philosophies toward model development: i) a direct maximization of business performance (GA), ii) a model selection approach, which introduces business objectives ex-post and develops models using statistical learning (Logit, RF, BBM and PAES), and iii) PCES that shifts the consideration of the actual business objective to a previous modeling stage to gear model development toward the ultimate goal of the marketing campaign.

We find the direct approach to be least effective. Even a simple logit model outperforms GA. The logit and GA model both construct a linear classifier. Better performance of the former evidences that model development through minimizing statistical loss is preferable to a direct maximization of business performance. Well-known estimation problems such as overfitting (Hastie et al., 2009) are a likely cause of this result. Remedies to such problems are available in statistical learning. However, developing predictive models through profit maximization, the direct approach is unable to capitalize on this knowledge.

Considering the model selection approach, logistic regression, random forest, and BBM perform better than GA but inferior to PCES. Profit improvements over these benchmarks are often

substantial. On average, PCES also recommends smaller campaigns, which indicates better targeting of PCES campaigns. Overall, these results suggest that incorporating business goals early in the modeling process has a sizeable positive effect on the quality of the prediction model and decision support, respectively.

One might object that a targeting model that is tuned to maximize profits will naturally give higher profits than a model that minimizes NLL or another loss function. Following this line of reasoning, one might question the fairness of the comparison in terms of campaign profit Eq. 3.1. However, it is important to recall that targeting is a prediction problem. We aim at predicting customer responses to marketing messages. In predictive modelling, it is crucial to develop a model on one set of (training) data and test it on a different, 'fresh' set of (test) data (Shmueli & Koppius, 2011). Given disjoint data sets for model training and evaluation, it is wrong to assume that maximizing profit on the training set will naturally give higher profit on the test set. This is apparent from the poor results of the GA benchmark and, more importantly, statistical learning theory (Vapnik & Kotz, 2006). Consequently, the experimental design facilitates a fair comparison.

However, it is still interesting to examine the performance of PCES across different evaluation measures to shed lights on the antecedents of its success in the above comparison. In particular, maximizing campaign profit Eq. 3.1 over $l(\tau)$ and τ, our evaluation criterion differs notably from typical accuracy indicators and statistical loss functions. We hypothesize that the advantage of PCES over benchmark models decreases when the ensemble selection criterion (i.e., business performance measure) is more similar to the loss functions that standard targeting models embody. To test this, the paper is accompanied by an e-companion, which provides results for additional performance measures; namely AUC and TDL (see Appendix 3.B) and campaign profit under a budget constraint (see Appendix 3.C). With respect to the similarity of these measures to standard indicators of predictive accuracy and statistical loss, we suggest an ordering of the form $AUC \prec TDL \prec \Omega\left(l\left(\tau\right), \tau = const.\right) \prec \Omega\left(l\left(\tau\right), \tau\right)$. AUC captures a classifier's ranking performance. It is a standard accuracy indicator, which we consider relatively closest to standard loss functions like NLL (Bequé et al., 2017). TDL is related to AUC but focuses on ranking performance among of subset of customers (Neslin et al., 2006). Thus, we consider it more distinct from model-internal loss functions. The same logic applies to campaign profit under a budget constrain ($\Omega\left(l\left(\tau\right), \tau = const.\right)$), just that this measure, in addition, depends on cost and benefit parameters which introduce further differences. Last, the evaluation measure we consider above, campaign profit with flexible marketing budget, $\Omega\left(l\left(\tau\right), \tau\right)$, includes the additional decision variable τ and is therefore most distinct from NLL or other standard loss functions.

Below, we summarize results from the e-companion and illustrate how the relative performance advantage of PCES develops across different performance measures. In particular, Table 3.6 reports the estimated performance improvement over a benchmark model across AUC, TDL, and campaign profit with fixed and flexible budget, whereby we use the same approach toward performance contrast estimation as in Table 3.4 (García et al., 2010). The e-companion provides

a more detailed analysis of AUC, TDL performance in Appendix 3.B, and campaign profit with budget constraint in Appendix 3.C.

Table 3.6: Comparison of PCES and benchmarks across statistical and monetary performance measures

	AUC	TDL	$\Omega\left(l\left(\tau\right),\tau = const.\right)$	$\Omega\left(l\left(\tau\right),\tau\right)$
Logit	7.31%	25.79%	18.10%	22.00%
RF	1.39%	3.58%	2.30%	14.00%
BBM	0.28%	3.10%	1.00%	7.00%
GA	6.23%	21.91%	15.60%	27.00%
PAES	0.00%	0.14%	0.30%	5.00%

We compute the relative performance improvements of PCES over benchmarks in the same way as in Table 3.4 using the contrast estimation approach of García et al. (2010).

Table 3.6 supports the view that PCES is most effective if an application specific (business) performance measure embodies a different notion of model performance than a model-internal loss function. Performance improvements are especially pronounced when assessing model performance in terms of campaign profit with flexible budget. On the other hand, improvements over the strongest competitor, PAES, vanish when using the AUC for performance evaluation, and are marginal for TDL and campaign profit under a budget constraint. The results for other benchmarks follow a similar trend, whereby PCES still provides a sizeable advantage in most cases. Overall, we take Table 3.6 as further evidence that incorporating profit consideration into model development is valuable. More specifically, the efficacy of PCES increases with decreasing similarity between a targeting model's internal loss function and a relevant measure of business performance.

3.7 Summary

We set out to develop a modeling approach that integrates principles of statistical learning with business objectives in customer targeting. To achieve this, we propose PCES, which first estimates a set of statistical prediction models and then selects from this library a subset of models so as to maximize campaign profit. The results that we obtain from a comprehensive empirical study confirm the effectiveness of this approach. We observe PCES to predict customer responsiveness more accurately than benchmarks and show that the profit of a marketing campaign increases when using PCES for target group selection. We also find this advantage over competitors to increase with decreasing correlation between a model-internal loss function and a relevant measure of business performance.

3.7.1 Implications

The results of our study have several implications. First, integrating business goals into the modeling process is interesting from a theoretical point of view. A large number of prediction methods have been developed in the literature. Well-grounded in the theory of statistical learning, such methods facilitate the development of empirical prediction models in diverse ap-

plication settings. Generality, however, has a cost. General purpose methods disregard the characteristic properties of specific applications such as profit in campaign planning. On the other end, a common approach toward decision support in the literature involves the development of tailor-made models that fully reflect the requirements of a given application. However, tailor-made models also suffer limitations. In the case of predictive modeling, a possible shortcoming may be that they are less accurate, for example because they fail to automatically account for nonlinear patterns. We consider our results a stimulus to rethink approaches to develop prediction models. In particular, we call for the development of modeling methodologies that are both widely applicable and aware of characteristic application requirements. To some extent, the proposed PCES framework is such an approach. For example, to adapt PCES to a decision problem other than targeting, we can replace the campaign profit function Eq. 3.1, which guides ensemble member selection, with an objective function that captures the peculiarities of the novel business application.

Second, from a managerial perspective, the key question is to what extent novel targeting models add to the bottom line. In this sense, an implication of our study is that it is feasible and effective to develop targeting models in a profit-conscious manner. Improvements of campaign profit of several percent, which we observe in many experimental settings, are managerially meaningful and indicate that PCES is a useful addition to campaign planners' toolset. Its application seems especially rewarding in settings where companies contact a large number of customers, conduct many campaigns, and/or run campaigns with high frequency, all of which is common in digital marketing and e-commerce.

A third implication of the study is related to the way in which targeting models are commonly employed in academia and industry. In particular, a model selection approach, which involves developing a set of candidate models and selecting *one* best model for deployment should be avoided. Our study suggests that an appropriately chosen combination of (some of these) alternative models using ensemble selection is likely to increase predictive accuracy and, more generally, model performance. Furthermore, introducing an additional selection and combination step into the modeling process provides an excellent opportunity to account for business objectives during model development.

Finally, a fourth implication is that the development of targeting models requires little human intervention. Typical modeling tasks include, for example, testing different variables, transformations of variables to increase their predictive value, and testing alternative prediction methods. Using an ensemble selection framework, campaign managers can easily automate these tasks. They only need to incorporate the candidate models that represent choice alternatives into the model library. The selection strategy will then pick the most beneficial model combination in a profit-conscious manner. This frees campaign planners from laborious, repetitive modeling tasks and unlocks valuable resources, which can be spend on tasks that truly require creativity and domain knowledge. In the case of predictive modeling, engineering informative features is a good example for such task.

3.7.2 Limitations and Future Research

Clearly, the study exhibits limitations that open up avenues for further research. Most importantly, we do not account for heterogeneity among customer values. We examine a range of settings in which the return per accepted offer differ. However, the return is always the same across customers. Given that customer spending differs in many practical applications, it is important to examine customer-dependent returns in future research. Future research could also extend the proposed modeling framework. In particular, PCES is a black-box approach that does not reveal how customer characteristics influence predictions. Such insight is important to understand which factors determine customers' reactions toward marketing offers. Therefore, developing approaches that unlock the PCES black-box and clarify how variables influence predictions seems to be a fruitful avenue for future research.

Finally, our study does not consider deep learning. This may seem surprising because deep learning methods have achieved excellent results, especially when processing unstructured data in computer vision and text analysis (LeCun et al., 2015). In this regard, examining the suitability of deep learning for customer targeting appears an interesting avenue for future research. However, the popular deep learning architectures convolutional and recurrent neural networks are particularly suitable for processing multi-dimensional data structures such as images (multiple images each of which consists of multiple pixels each of which has multiple color channels) or texts (multiple documents each of which consists of multiple words, each of which is projected to a multi-dimensional embedding space), and appear less appropriate for the cross-sectional data we employ here and that prevails in the literature on customer targeting (Martens et al., 2016). For example, tabular data with two dimensions, observations and features, does not exhibit a sequential structure, which discourages application of recurrent networks. Similarly, the filtering operation in convolutional networks is not readily applicable when working with "flat tables". In view of this, future research on deep learning-based targeting would benefit from multi-dimensional input data where the values of individual features are available over time. Until corresponding results become available, interested readers find a preliminary analysis of a deep network with our data sets in the online appendix that accompanies this paper. For these data sets, we find PCES to perform significantly and substantially better than a deep learning benchmark.

Bibliography

Bequé, A., Coussement, K., Gayler, R., & Lessmann, S. (2017). Approaches for credit scorecard calibration: An empirical analysis. *Knowledge-Based Systems*, *134*(15), 213–227. https://doi.org/10.1016/j.knosys.2017.07.034

Bhattacharyya, S. (1999). Direct marketing performance modeling using genetic algorithms. *INFORMS Journal on Computing*, *11*(3), 248–257.

Blattberg, R. C., Neslin, S. A., & Kim, B.-D. (2008). *Database Marketing: Analyzing and Managing Customers*. New York, Springer.

Caruana, R., Munson, A., & Niculescu-Mizil, A. (2006). Getting the Most Out of Ensemble Selection, In *Proceedings of the 6th International Conference on Data Mining (ICDM'06)*, Los Alamitos, IEEE Computer Society. https://doi.org/10.1109/ICDM.2006.76

Christoffersen, P. F., & Diebold, F. X. (1997). Optimal prediction under asymmetric loss. *Econometric Theory*, *13*(6), 808–817.

Coussement, K., & Buckinx, W. (2011). A probability-mapping algorithm for calibrating the posterior probabilities: A direct marketing application. *European Journal of Operational Research*, *214*(3), 732–738. https://doi.org/10.1016/j.ejor.2011.05.027

Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems*, *95*, 27–36. https://doi.org/10.1016/j.dss.2016.11.007

Cui, G., Wong, M. L., & Wan, X. (2015). Targeting high value customers while under resource constraint: Partial order constrained optimization with Genetic Algorithm. *Journal of Interactive Marketing*, *29*, 27–37. https://doi.org/http://dx.doi.org/10.1016/j.intmar.2014.09.001

Ding, A. W., Li, S., & Chatterjee, P. (2015). Learning user real-time intent for optimal dynamic web page transformation. *Information Systems Research*, *26*(2), 339–359. https://doi.org/10.1287/isre.2015.0568

Elsner, R., Krafft, M., & Huchzermeier, A. (2004). Optimizing Rhenania's direct marketing business through dynamic multilevel modeling (DMLM) in a multicatalog-brand environment. *Marketing Science*, *23*(2), 192–206.

Fuller, R. M., & Dennis, A. R. (2009). Does fit matter? The impact of task-technology fit and appropriation on team performance in repeated tasks. *Information Systems Research*, *20*(1), 2–17. https://doi.org/10.1287/isre.1070.0167

García, S., Fernández, A., Luengo, J., & Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, *180*(10), 2044–2064.

Geng, Y., Liang, R., Li, W., Wang, J., Liang, G., Xu, C., & Wang, J. (2016). Learning Convolutional Neural Network to Maximize Pos@Top Performance Measure, In *Proceedings of the European Symposium on Artificial Neural Networks (ESANN'2016)*.

Glady, N., Baesens, B., & Croux, C. (2009). Modeling churn using customer lifetime value. *European Journal of Operational Research*, *197*(1), 402–411.

Golrezaei, N., Nazerzadeh, H., & Rusmevichientong, P. (2014). Real-time optimization of personalized assortments. *Management Science*, *60*(6), 1532–1551. https://doi.org/10.1287/mnsc.2014.1939

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Granger, C. W. J. (1969). Prediction with a generalized cost of error function. *Operational Research Quarterly*, *20*(2), 199–207.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The Elements of Statistical Learning* (Second). New York, Springer.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: Online Appendix. *European Journal of Operational Research*, *247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Li, C.-T., Lin, Y.-J., & Yeh, M.-Y. (2018). Forecasting participants of information diffusion on social networks with its applications. *Information Sciences*, *422*, 432–446. https://doi.org/10.1016/j.ins.2017.09.034

Lichman, M. (2013). UCI Machine Learning Repository.

Lilien, G. L. (2011). Bridging the academic–practitioner divide in marketing decision models. *Journal of Marketing*, *75*(4), 196–210.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, *234*, 11–26. https://doi.org/10.1016/j.neucom.2016.12.038

Maldonado, S., Bravo, C., López, J., & Pérez, J. (2017). Integrated framework for profit-based feature selection and SVM classification in credit scoring. *Decision Support Systems*, *104*, 113–121. https://doi.org/10.1016/j.dss.2017.10.007

Maldonado, S., Pérez, J., & Bravo, C. (2017). Cost-based feature selection for Support Vector Machines: An application in credit scoring. *European Journal of Operational Research*, *261*(2), 656–665. https://doi.org/10.1016/j.ejor.2017.02.037

Martens, D., & Provost, F. (2011). *Pseudo-Social Network Targeting from Consumer Transaction Data* (tech. rep.). NYU Working Paper No. CEDER-11-05.

Martens, D., Provost, F., Clark, J., & Junqué de Fortuny, E. (2016). Mining massive fine-grained behavior data to improve predictive analytics. *MIS Quarterly*, *40*(4), 869–888. https://doi.org/10.25300/misq/2016/40.4.04

Mitrović, S., Baesens, B., Lemahieu, W., & De Weerdt, J. (2018). On the operational efficiency of different feature types for telco churn prediction. *European Journal of Operational Research*, *267*(3), 1141–1155. https://doi.org/10.1016/j.ejor.2017.12.015

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research*, *43*(2), 204–211. https://doi.org/10.1509/jmkr.43.2.204

Perlich, C., Dalessandro, B., Raeder, T., Stitelman, O., & Provost, F. (2014). Machine learning for targeted display advertising: Transfer learning in action. *Machine Learning*, *95*(1), 103–127. https://doi.org/10.1007/s10994-013-5375-2

Platt, J. C. (2000). Probabilities for Support Vector Machines. In A. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), *Advances in Large Margin Classifiers* (pp. 61–74). Cambridge, MIT Press.

Rokach, L., Naamani, L., & Shmilovici, A. (2008). Pessimistic cost-sensitive active learning of decision trees for profit maximizing targeting campaigns. *Data Mining and Knowledge Discovery*, *17*(2), 283–316.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*(0), 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, *35*(3), 553–572.

Spanoudes, P., & Nguyen, T. (2017). Deep learning in customer churn prediction: Unsupervised feature learning on abstract company independent feature vectors. *arXiv preprint*, arXiv:1703.03869v1.

Stripling, E., vanden Broucke, S., Antonio, K., Baesens, B., & Snoeck, M. (2018). Profit maximizing logistic model for customer churn prediction using genetic algorithms. *Swarm and Evolutionary Computation*, *40*, 116–130.

Tambe, P. (2014). Big data investment, skills, and firm value. *Management Science*, *60*(6), 1452–1469. https://doi.org/doi:10.1287/mnsc.2014.1899

Van Vlasselaer, V., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2017). GOTCHA! Network-based fraud detection for social security fraud. *Management Science*, *63*(9), 3090–3110. https://doi.org/10.1287/mnsc.2016.2489

Vapnik, V., & Kotz, S. (2006). *Estimation of Dependences Based on Empirical Data* (Second). New York, Springer.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, *218*(1), 211–229. https://doi.org/10.1016/j.ejor.2011.09.031

Verbraken, T., Bravo, C., Weber, R., & Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research*, *238*(2), 505–513. https://doi.org/10.1016/j.ejor.2014.04.001

Verbraken, T., Verbeke, W., & Baesens, B. (2012). A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE Transactions on Knowledge and Data Engineering*, *25*(5), 961–973. https://doi.org/10.1109/TKDE.2012.50

Wang, S., Li, Y., Shao, Y., Cattani, C., Zhang, Y., & Du, S. (2017). Detection of dendritic spines using wavelet packet entropy and fuzzy support vector machine. *CNS & Neurological Disorders - Drug Targets*, *16*(2), 116–121.

Zhao, H., & Li, X. (2017). A cost sensitive decision tree algorithm based on weighted class distribution with batch deleting attribute mechanism. *Information Sciences*, *378*, 303–316. https://doi.org/10.1016/j.ins.2016.09.054

Zhu, B., Baesens, B., & vanden Broucke, S. K. L. M. (2017). An empirical comparison of techniques for the class imbalance problem in churn prediction. *Information Sciences*, *408*, 84–99. https://doi.org/10.1016/j.ins.2017.04.015

3.A Working Example of Ensemble Selection

Table 3.7 illustrates the member selection strategy within ensemble selection using a synthetic example. Suppose we have five customer records, a binary response variable y, and a library of four candidate models (M1–M4). Each model estimates the probability that a customer accepts a marketing offer (y=1) or refuses to do so (y=0). For simplicity, we assess predictions in terms

of mean squared error (MSE). M4 is the best individual model (lowest MSE). We thus select M4 as first ensemble member (iteration 1). In the second iteration, we assess all size-two ensembles (including M4) by combining models through averaging their forecasts (depicted by an "&" in Table 3.7). Given that the M4 & M1 ensemble achieves better performance than M4 alone (MSE .26, compared to .27), we append M1 to the ensemble and continue the selection process. Since the third iteration achieves no further error reduction, we stop ensemble growing. The final ensemble is given by the combination of M4 and M1.

Table 3.7: Illustration of ensemble selection on MSE with a library of four candidate models

Iteration 1	**M1**	**M2**	**M3**	**M4**	***y***
	0.89	0.33	0.43	0.45	1
	0.87	0.84	0.64	0.55	0
	0.31	0.37	0.90	0.69	0
	0.49	0.83	0.69	0.70	1
	0.65	0.85	0.60	0.38	0
MSE	0.31	0.41	0.40	**0.27**	
Iteration 2	**M1 & M4**	**M2 & M4**	**M3 & M4**	**M4 & M4**	y
	0.67	0.39	0.44	0.45	1
	0.71	0.69	0.59	0.55	0
	0.50	0.53	0.80	0.69	0
	0.60	0.76	0.70	0.70	1
	0.52	0.62	0.49	0.38	0
MSE	**0.26**	0.32	0.33	0.27	
Iteration 3	**M1 & M4 & M1**	**M2 & M4 & M1**	**M3 & M4 & M1**	**M4 & M4 & M1**	y
	0.74	0.55	0.59	0.59	1
	0.76	0.75	0.69	0.66	0
	0.44	0.46	0.64	0.57	0
	0.56	0.67	0.63	0.63	1
	0.56	0.63	0.55	0.47	0
MSE	0.27	0.30	0.30	**0.26**	

Bold font highlights the best performing candidate ensemble per iteration by minimum MSE

3.B Statistical Comparison of Targeting Models

In order to complement the results presented in the main part of the paper, Appendix 3.B elaborates on comparative results in terms of indicators that are widely used in classifier benchmarks and marketing applications: the area under the receiver operating characteristics curve (AUC) and top-decile lift (TDL).

The AUC is equivalent to the Gini coefficient and captures the degree to which a targeting model succeeds in estimating higher response probabilities to actual campaign respondents (Neslin et al., 2006). Table 3.8 depicts AUC results for PCES and benchmark models. In addition, the second to last row of Table 3.8 reports the contrast between PCES and a benchmark, which we estimate on the basis of the median AUC difference (García et al., 2010). We also compare the statistical significance of AUC differences using the Friedman test. For the results of Table 3.8, a χ^2 value of 90.5 indicates that we can reject the null hypothesis of equal performance (p-value $<$0.000). This facilitates proceeding with multiple pairwise comparisons of PCES against one

benchmark. The last row of Table 3.8 reports the corresponding p-values, which we adjust using Rom's procedure (García et al., 2010).

Table 3.8 suggests that PCES compares favorably against the benchmarks, yielding the highest AUC value on 16 out of 25 data sets. Given these results, we may conclude that PCES predicts significantly more accurately (in terms of the AUC) than all benchmarks but PAES. In case of the latter, empirical evidence does not suffice to reject the null-hypothesis of equal performance. Considering the magnitude of performance differences, the second to last row of Table 3.8 indicates that performance differences in terms of the AUC are sizeable for weaker benchmarks (e.g., around 0.04 AUC points compared to logit and GA) but marginal against state-of-the-art benchmarks such as RF and BBM. This reemphasizes the differences in terms of predictive accuracy and business performance. Put differently, the results of Table 3.8 add to the main part of the paper in that they show how small differences in terms of the AUC can translate into meaningful differences in campaign profit, which we observe in the main part of the paper. Last, performance differences in terms of TDL, which we report in Table 3.9, mimic the AUC comparison to large extent. For example, PCES performs significantly better than the Logit, RF, BBM, and GA benchmark and competitive to PAES.

The empirical results presented above provide strong evidence of the predictive accuracy of PCES. PCES gives higher AUC and TDL than benchmark models and statistical tests confirm that PCES performs significantly better than these benchmarks. However, one may object that the benchmarks do not consider one of the latest developments in machine learning. As acknowledged in the main part of the paper, we do not include prediction models based on deep learning algorithms in our model libraries. Deep learning is a broad field that has given rise to many exciting modeling techniques and network architectures (LeCun et al., 2015; Liu et al., 2017; Schmidhuber, 2015). In judging the potential of corresponding techniques for customer targeting, however, it is important to note that some of these architectures are inapplicable or less suitable for the data sets we employ here. In particular, recurrent neural networks are designed to process sequential data while we work with cross-sectional data. Our data represents a snapshot that has been gathered at a given point in time and does thus not embody temporal information. Similarly, convolutional neural networks (CNNs), which are popular for text and image analysis, make use of filters that slide through the input data to detect patterns (Goodfellow et al., 2016). In image analysis, for example, the sliding filter operation processes neighboring pixels. In text analytics, a filter slides over neighboring words in a sentence. This operation is undefined in cross-sectional data that consists of a flat table with two dimensions of observations and features. In theory, a filter could process neighboring observations in such a data set. However, the order of observations in a cross-sectional data is random. Alternatively, a filter could process neighboring features. Yet, the order in which features appear in a cross-sectional data set is arbitrary. Consequently, the filtering mechanism, which is instrumental to CNNs, cannot detect meaningful patterns in cross-sectional data.

Deep learning architectures that are more suitable for our cross-sectional data include stacked auto-encoders and deep believe networks, both of which perform unsupervised pre-training,

Table 3.8: Comparison of predictive performance in terms of the AUC

Data set	**Logit**	**RF**	**BBM**	**GA**	**PAES**	**PCES**
D1	0,637	0,651	0,665	0,656	**0,676**	0,674
D2	*0,914*	*0,998*	*0,998*	*0,945*	*0,999*	**0,999**
D3	0,641	0,650	0,650	0,647	**0,674**	0,667
D4	*0,569*	0,695	0,695	*0,603*	**0,697**	0,695
D5	*0,586*	*0,688*	*0,688*	*0,607*	**0,691**	0,691
D6	*0,768*	*0,822*	*0,832*	*0,774*	*0,836*	**0,842**
D7	*0,584*	*0,669*	*0,669*	*0,609*	*0,677*	**0,678**
D8	*0,670*	*0,665*	*0,690*	*0,656*	*0,697*	**0,698**
D9	*0,613*	*0,611*	0,631	*0,586*	**0,632**	0,631
D10	*0,702*	*0,736*	*0,763*	*0,742*	*0,768*	**0,769**
D11	*0,798*	*0,907*	*0,906*	*0,815*	**0,911**	0,908
D12	*0,578*	*0,589*	*0,592*	*0,573*	*0,593*	**0,594**
D13	*0,562*	*0,785*	*0,764*	*0,622*	*0,795*	**0,795**
D14	*0,894*	*0,917*	*0,926*	*0,903*	*0,928*	**0,928**
D15	*0,808*	0,852	0,857	0,848	**0,857**	0,842
D16	*0,625*	*0,588*	*0,625*	*0,606*	*0,628*	**0,629**
D17	**0,757**	0,753	0,753	*0,698*	*0,745*	0,752
D18	*0,802*	*0,958*	*0,963*	*0,787*	*0,971*	**0,973**
D19	*0,857*	*0,963*	*0,968*	*0,847*	*0,967*	**0,970**
D20	*0,628*	*0,664*	*0,672*	*0,639*	*0,675*	**0,677**
D21	*0,898*	*0,930*	*0,930*	*0,893*	*0,932*	**0,934**
D22	*0,705*	*0,722*	*0,721*	*0,706*	*0,724*	**0,726**
D23	*0,658*	*0,686*	*0,699*	*0,658*	*0,699*	**0,699**
D24	*0,533*	0,603	**0,609**	*0,601*	0,608	0,602
D25	*0,673*	*0,742*	*0,745*	*0,693*	*0,747*	**0,747**
Contrast	0,049	0,010	0,002	0,041	0,000	
p-values	0,000	0,000	0,020	0,000	0,705	

Bold face highlights the best performing model per data set (highest AUC). In addition, italic font indicates that a model performs worse than PCES. Note that the formatting is based on the exact AUC results, whereas Table 3.8 shows AUC values rounded to three digits of accuracy. The computation of performance contrasts and the adjustment of p-values in multiple pairwise classifier comparisons are based on García et al. (2010).

Table 3.9: Comparison of predictive performance in terms of TDL

Data set	**Logit**	**RF**	**BBM**	**GA**	**PAES**	**PCES**
D1	*1,743*	*2,142*	*2,017*	*1,619*	**2,229**	**2,229**
D2	*5,236*	*7,123*	*7,123*	*6,646*	*7,121*	**7,127**
D3	*1,756*	*1,812*	*1,812*	*1,999*	**2,241**	2,148
D4	*1,320*	2,765	2,765	*1,885*	**2,828**	2,702
D5	*1,701*	*2,711*	*2,711*	*1,887*	**3,057**	3,004
D6	*4,119*	*5,228*	*5,545*	*4,436*	**5,862**	5,703
D7	*1,374*	*1,800*	*1,800*	*1,419*	*1,863*	**1,873**
D8	*2,267*	*2,355*	*2,453*	*2,091*	*2,707*	**2,717**
D9	*1,857*	*1,796*	**1,967**	*1,487*	1,953	1,933
D10	*3,142*	*3,559*	*3,810*	*3,281*	*3,782*	**3,935**
D11	*3,179*	*6,786*	*6,786*	*4,321*	*6,857*	**6,893**
D12	*1,678*	*2,007*	*1,981*	*1,739*	**2,042**	2,024
D13	*1,603*	**1,984**	*1,930*	*1,754*	**1,984**	1,978
D14	*3,498*	*3,838*	3,930	*3,584*	**3,935**	3,922
D15	*2,715*	*3,374*	*3,386*	*3,121*	*3,395*	**3,404**
D16	*1,952*	*1,673*	*1,923*	*1,673*	*2,055*	**2,073**
D17	*3,900*	*3,456*	*3,456*	*3,626*	*3,902*	**3,941**
D18	*4,044*	*7,034*	*7,182*	*4,071*	*7,667*	**7,690**
D19	*5,253*	*8,251*	*8,471*	*5,083*	**8,543**	8,543
D20	*1,962*	*2,106*	*2,050*	*1,862*	*2,142*	**2,154**
D21	*5,034*	*5,251*	*5,222*	*4,784*	*5,303*	**5,350**
D22	*2,769*	**3,288**	*2,899*	*2,856*	*2,985*	3,159
D23	*1,280*	*1,381*	1,449	*1,280*	**1,458**	1,445
D24	*1,324*	*1,779*	*1,836*	*1,529*	**1,894**	1,868
D25	*1,893*	*2,473*	**2,481**	*2,017*	2,480	2,477
Contrast	0,506	0,097	0,084	0,442	0,004	
p-values	0,000	0,015	0,015	0,000	0,910	

Bold face highlights the best performing model per data set (highest TDL). In addition, italic font indicates that a model performs worse than PCES. Note that the formatting is based on the exact TDL results, whereas Table 3.9 shows TDL values rounded to three digits of accuracy. The computation of performance contrasts and the adjustment of p-values in multiple pairwise classifier comparisons are based on García et al. (2010).

and deep feed-forward neural networks (DFFNN), which can be considered a generalization of the 'shallow' single hidden layer neural networks we consider in our model libraries (see Table 3.1). According to Goodfellow et al. (2016), unsupervised pre-training is today mainly used for natural language processing (Goodfellow et al., 2016, p.526). The authors are also critical with deep belief networks (Goodfellow et al., 2016, p.651ff). These recommendations from leading deep learning experts, together with the inapplicability of CNNs and recurrent networks suggest that DFFNN are a competitive deep learning approach for cross-sectional data. This approach is also considered in Spanoudes and Nguyen (2017), which is one of the very few papers on deep learning-based prediction in marketing. Therefore, to support the results presented above and in the main part of the paper, we compare PCES to a DFFNN in terms of AUC and TDL in Table 3.10. We use the same approach for statistical testing and format results in the same way as elsewhere in the paper and online appendix. To offer an additional context for the interpretation of empirical results, we also include the BBM benchmark in the comparison.

Table 3.10: Comparison of PCES to a deep feedforward neural network (DFFNN)

Data set	**AUC**			**TDL**		
	DFFNN	**BBM**	**PCES**	**DFFNN**	**BBM**	**PCES**
D1	0,655	0,665	**0,674**	2,027	2,017	**2,229**
D2	0,994	0,998	**0,999**	7,091	7,123	**7,127**
D3	0,487	0,65	**0,667**	0,765	1,812	**2,148**
D4	0,620	**0,695**	**0,695**	1,323	**2,765**	2,702
D5	0,650	0,688	**0,691**	1,473	2,711	**3,004**
D6	0,818	0,832	**0,842**	2,429	5,545	**5,703**
D7	0,639	0,669	**0,678**	1,622	1,800	**1,873**
D8	0,580	0,690	**0,698**	1,815	2,453	**2,717**
D9	0,544	**0,631**	**0,631**	1,432	**1,967**	1,933
D10	0,751	0,763	**0,769**	3,613	3,810	**3,935**
D11	0,892	0,906	**0,908**	5,615	6,786	**6,893**
D12	0,570	0,592	**0,594**	1,764	1,981	**2,024**
D13	0,765	0,764	**0,795**	1,971	1,930	**1,978**
D14	0,909	0,926	**0,928**	3,619	**3,930**	3,922
D15	0,851	**0,857**	0,842	3,311	3,386	**3,404**
D16	0,623	0,625	**0,629**	1,918	1,923	**2,073**
D17	0,699	**0,753**	0,752	3,510	3,456	**3,941**
D18	**0,980**	0,963	0,973	7,499	7,182	**7,69**
D19	0,952	0,968	**0,970**	7,742	8,471	**8,543**
D20	0,670	0,672	**0,677**	2,072	2,050	**2,154**
D21	0,922	0,930	**0,934**	5,248	5,222	**5,350**
D22	0,703	0,721	**0,726**	**3,237**	2,899	3,159
D23	0,673	**0,699**	**0,699**	1,318	**1,449**	1,445
D24	0,589	**0,609**	0,602	1,414	1,836	**1,868**
D25	0,746	0,745	**0,747**	2,475	**2,481**	2,477
Contrast	0,019	0,003		0,267	0,0967	
p-values	0,000	0,016		0,000	0,003	

Bold face highlights the best performing model per data set (highest AUC or TDL). Note that the formatting is based on the exact results of the AUC and TDL statistic, whereas Table 3.10 shows AUC and TDL values rounded to three digits of accuracy. The computation of performance contrasts and the adjustment of p-values in multiple pairwise classifier comparisons are based on García et al. (2010).

Table 3.10 further supports the view that PCES is a powerful approach for prediction. On the

majority of data sets, PCES-based AUC and TDL statistics are higher compared to the BBM and DFFNN benchmark and statistical tests suggest that PCES performs significantly better than these benchmarks. The performance of the DFFNN model is disappointing. We observe only one case where this approach provides the highest AUC value. Similarly, DFFNN gives the largest TDL result on one data set. In fact, the DFFNN model performs not only inferior to PCES but also loses in a direct comparison to the BBM benchmark. One the one hand, poor performance of this specific deep learning model might seem surprising, given that deep learning has given excellent results in a number of applications (LeCun et al., 2015). One the other hand, the particular type of data we employ in this paper is arguably less complex than, e.g., natural language, high-dimensional, nonstationary temporal data, or, more generally, the types of data and applications where deep learning is predominantly considered. Table 3.10 suggests that conventional machine learning methods, which are underneath the BBM benchmark (see Table 3.1), perform competitive or even better than the DFFNN on the cross-sectional data sets used in this study. Of course, this finding is only valid for the specific data sets considered in the paper. Due to the considerable scope of the comparison (i.e., 25 marketing data sets) results of Table 3.10 may be considered relevant evidence for the task of customer targeting, while a further generalization of observed results is clearly inappropriate. In this regard, we strongly encourage future research to test deep learning in other marketing tasks and uncover the antecedents of its success or failure. As far as this paper is concerned, Table 3.10 supports the conclusion that the sparse coverage of deep learning in our analysis does not put the appealing performance of the proposed PCES approach into perspective. This follows directly from the strong superiority of PCES over the DFFNN model in Table 3.10, which can be considered a representative benchmark of the deep learning approaches that are applicable to the data sets under study. To further support this view, it is important to note that the results of Table 3.10 estimate the performance of PCES in a conservative manner because the DFFNN model has not been added to the model library. That is, we do not rebuild our PCES models for Table 3.10. As a consequence, PCES does not have access to the forecasts of the DFFNN model and cannot incorporate DFFNN models in the ensemble to further raise predictive accuracy. Such incorporation of the DFFNN model in the ensemble is likely in cases such as that of data set D18 or D22 where the DFFNN model performs well.

3.C Campaign Profit Maximization Under a Budget Constraint

In this setting, we assume that a marketer strives to maximize campaign profit given a fixed budget. This case can arise if the responsibility for the overall campaign is with the marketing department, which decides on budgets, whereas an analytics unit or external contractor is in charge of the actual targeting. A fixed marketing budget implies a fixed campaign size, where τ is chosen such that the budget is exhausted. With τ externally given, maximizing campaign profit (1) is equivalent to maximizing lift. To increase generality, we consider a range of different campaign sizes and select τ from the interval 0.05, 0.10, ..., 0.45. This way, we obtain 9 (campaign sizes) $\times$ 25 (data sets) = 225 estimates of campaign profit per targeting model. Table 3.11 summarizes these results in the form of win-tie-loss statistics. As in Table 3 in the main part of the paper, we compare the statistical significance of profit differences using

the Friedman test (bottom of Table 3.11) and reject the null hypothesis of equal performance ($\chi^2 = 710.6$, p-value <0.000). This allows us to proceed with pairwise comparisons of PCES against one benchmark to detect significant differences among individual targeting models. To protect against an elevation of alpha values in multiple pairwise comparisons, we adjust p-values using Rom's procedure (García et al., 2010). The last row of Table 3.11 reports the adjusted p-values.

Table 3.11: Win-tie-loss statistics of PCES versus benchmarks for fixed campaign sizes

Campaign	PCES vs. Logit			PCES vs. RF			PCES vs. BBM			PCES vs. GA			PCES vs. PAES		
size (τ)	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss
0.05	24	1	0	20	4	1	17	6	2	24	0	1	12	8	5
0.10	25	0	0	22	0	3	20	0	5	25	0	0	12	1	12
0.15	24	0	1	22	1	2	22	2	1	25	0	0	13	2	10
0.20	24	0	1	21	1	3	21	1	3	25	0	0	16	2	7
0.25	25	0	0	23	0	2	20	2	3	24	0	1	17	1	7
0.30	21	0	4	19	2	4	17	3	5	23	0	2	12	4	9
0.35	22	0	3	18	2	5	14	3	8	24	0	1	13	2	10
0.40	24	0	1	16	1	8	15	1	9	24	0	1	15	3	7
0.45	24	0	1	18	1	6	18	3	4	25	0	0	13	1	11
Total	213	1	11	179	12	34	164	21	40	219	0	6	123	24	78
	95%	0%	5%	80%	5%	15%	73%	9%	18%	97%	0%	3%	55%	11%	35%
p-value		0.000			0.000			0.000			0.000			0.123	

Note that results in the row τ=0.10 correspond to the popular top-decile-lift. The p-values refer to pairwise comparisons of PCES versus one benchmark, and are adjusted using Rom's procedure (García et al., 2010).

Table 3.11 evidences that PCES performs significantly better than the logit, RF, and BBM benchmarks. Specifically, PCES gives higher profit in 95, 80, and 73 percent of cases, respectively (p-values consistently less than 0.000). These results reemphasize that introducing the relevant notion of model performance during model development (as opposed to model selection) increases performance. In comparison to PAES, however, we find PCES to win, tie, and lose in 123 (55 percent), 24 (eleven percent), 78 (35 percent) pairwise comparisons. Statistical testing suggests this evidence to be insufficient to reject the null hypothesis of zero profit differences (p-value: 0.123). Before examining the relative performance of alternative targeting models in more detail, we note that the GA benchmark performs much worse than PCES.

To obtain a clearer view on the degree to which PCES increases business performance, we once again consider a fictitious company with a customer base of N = 100,000 customers and let the per-customer return from accepted offers, r, and offer costs to contact customers, c, be \$10 and \$1, respectively. These are the same settings as in the main part of the paper. Similarly, we re-use the above settings of $\tau = 0.05, 0.10, \ldots, 0.45$. Then, for each targeting model and campaign size, we estimate $l(\tau)$ through the median lift across our data sets (see main part of the paper as well as García et al. (2010)). Table 3.12 displays the total campaign profits that emerge from contacting the target group that each model selects for solicitation, together with the median percentage improvement of PCES over benchmark models.

Table 3.12 shows that using the logit model for targeting entails substantial opportunity costs. Compared to logit, PCES produces higher campaign profits across all τ settings and can be

Table 3.12: Campaign profit from different models for a fictitious marketing campaign

	Campaign profit $					
Campaign size (τ)	Logit	RF	BBM	GA	PAES	PCES
0.05	19,418	29,400	29,825	21,594	**30,051**	**30,051**
0.10	34,500	49,343	**49,543**	35,744	49,518	49,443
0.15	44,243	64,701	**65,570**	44,972	64,970	65,245
0.20	54,500	73,771	73,771	56,428	**76,092**	75,475
0.25	59,500	82,768	82,768	56,719	82,711	**83,391**
0.30	61,493	**83,018**	**83,018**	58,460	82,565	**83,018**
0.35	62,254	80,628	80,628	58,855	80,006	**80,968**
0.40	61,964	77,445	77,445	64,123	77,162	**77,502**
0.45	64,995	76,565	77,327	66,061	78,240	**78,392**
Profit increase using PCES*	18.1%	2.3%	1.0%	15.6%	0.3%	–

* We derive these values as follows: we first compute performance contrasts between targeting models for each campaign size (García et al., 2010), then convert the contrast to percentages by dividing through campaign profits (per campaign size), and finally calculate the median percentage increase over campaign sizes.

expected to increase profits by 18.1 percent on average. The RF and BBM targeting models represent more advanced benchmarks. Accordingly, PCES-induced profit increases are smaller and amount to 2.3 and 1.0 percent on average. A two percent profit increase against RF, a classifier much credited for high accuracy (Lessmann et al., 2015; Verbeke et al., 2012), is a sizeable improvement. Even a one percent improvement might be managerial meaningful, for example, for larger companies and campaigns.

Table 3.12 also reemphasizes the strengths of the PAES benchmark. Although PCES gives higher profits for seven of nine τ settings, the average profit increase is only 0.3%. Since PAES and PCES differ in terms of their target function, this difference is a result of our proposition to select ensemble members according to business objectives as opposed to statistical loss functions. Even if 0.3 percent can translate into large absolute monetary values, it is fair to question the business value of such small improvement. However, in appraising the PAES results, it is important to note that PAES and PCES entail the same effort. The computational cost of developing a targeting model with either one of these approaches lies in the construction of the base model library. In comparison, efforts concerned with base model selection using NLL (PAES) or campaign profit (PCES) are negligible. In fact, this is also true for the BBM benchmark, which we select as the strongest targeting model out of a large library of 887 candidate models (see Table 3.1). In that sense, Table 3.12 provides solid evidence that PCES is an effective targeting approach: it consistently performs as good as or better than benchmarks that entail comparable effort (BBM, PAES) and provides sizeable profit increases over computationally cheaper benchmarks (Logit, RF).

Last, we note that GA performs slightly better than the logit model in this comparison. Given that the logit model benefits from stronger theoretical underpinnings (Vapnik & Kotz, 2006), it is surprising to observe inferior results compared to the profit maximizing GA model. Recall

that GA is the weakest model in the analysis of campaign profit without budget constraint (i.e., main part of the paper). While a detailed analysis of this phenomenon is beyond the scope of this study, we take the good performance of GA compared to logit as evidence that the configuration of the GA model (meta-parameters, fitness function, use of GA as opposed to another search strategy) produces a competitive benchmark. Nonetheless, Table 3.12 further supports the conclusion from the main part of the paper that a completely profit-driven benchmark is inferior to PCES, which integrates statistical learning principles through the candidate library and economic consideration through the profit-maximizing ensemble selection.

Since the results of Table 3.12 are based on a specific choice of campaign parameters (r, c, N), we next repeat the sensitivity analysis (see main part of the paper) and examine the development of campaign profit across settings of r = \$2, \$3, \$5, \$10, \$15, \$25, \$50, \$75, and \$100. Figure 3.3 summarizes the empirical results of the 9 (settings for r) $\times$ 9 (settings for τ) $\times$ 25 (data sets) = 2025 comparisons between PCES and its benchmarks. In particular, it depicts the expected percentage increase in campaign profits that result from selecting target groups by means of PCES compared to a benchmark. We estimate profit increases in a similar way as in Table 3.12, but calculate the median of the difference between performance contrasts (García et al., 2010) across data sets and settings of r.

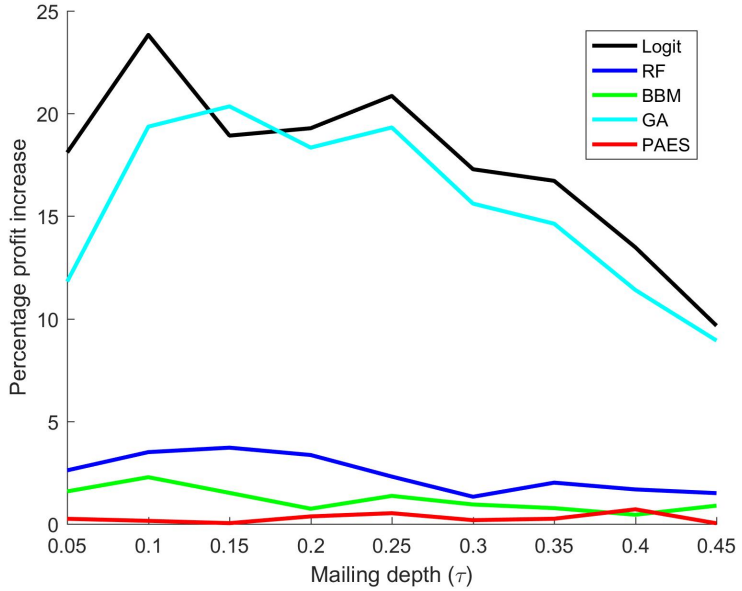


Figure 3.3: Expected percentage improvement in campaign profit due to using PCES for target group selection in a fixed budget setting. We estimate profit contrasts in the same way as in Table 3.12.

The main conclusion emerging from Figure 3.3 is that the interpretation of Table 3.12 generalizes to other settings of campaign parameters. Non-zero percentage profit differences for all comparisons show that PCES performs at least as good as and typically better than the benchmarks. More specifically, aggregating results over 25 marketing datasets and 9 settings of campaign parameters (r), we observe PCES to outperform the two linear benchmarks, logit and GA, with substantial margin. Compared to the more challenging RF and BBM benchmark, PCES increases campaign profit on average by one to four percent. Across all campaign sizes,

PAES is the most challenging benchmark. With one exception, where both methods tie, PCES achieves a small profit increase over PAES.

Figure 3.3 also suggests that profit differences between targeting models decrease with campaign size. A negative trend, which we observe for all benchmarks but PAES, is plausible because larger campaigns will inevitably contact more customers who do not accept marketing offers. In fact, the purpose of using a targeting model is to identify a — often small — subgroup of responsive customers (Blattberg et al., 2008). Therefore, we expect smaller settings of τ to represent a larger number of corporate applications and corresponding results especially relevant from a managerial point of view. It is encouraging to observe PCES to outperform several benchmarks with substantial margin at small campaign sizes of τ=0.05, 0.10, and 0.15.

Chapter 4

Revenue Uplift Modeling

Publication
Gubela, R. M., Lessmann, S., Haupt, J., Baumann, A., Radmer, T., & Gebert, F. (2017). Revenue Uplift Modeling. Proceedings of the 38th International Conference on Information Systems (ICIS).

Abstract
The measurement of the effectiveness of a marketing campaign is a challenging task. Whereas established approaches do not consider causality, uplift models take into account which customers display some behavior because of the marketing action and model this target as differential response. The paper categorizes existing approaches toward uplift modeling collected from different fields into a conceptual taxonomy to establish the state-of-the-art and proposes a novel approach named revenue uplift modeling. Contrary to existing approaches, which model incremental response, revenue uplift models predict the incremental revenue with the goal to maximize the gain per marketing incentive for heterogeneous customers. An experiment based on a large real-world dataset of e-commerce shops across several industries provides a benchmark on the choice of machine learning methods to implement the identified uplift modeling approaches and demonstrates the effectiveness of the revenue uplift model in a real-world e-commerce environment.

4.1 Introduction

Advertisements are omnipresent. A recent study of media use and advertisement exposure points out that the typical U.S. adult encounters a total of about 153 advertisements each day (Media Dynamic, Inc., 2014). Accordingly, advertising investment is substantial. In 2015 alone, about 161 billion USD were spent on digital advertising across all Internet-connected devices worldwide (eMarketer, 2016). To ensure accountability of investments and allocate marketing resources efficiently, it is important to measure the effectiveness of advertisement and more generally marketing communication. This remains a challenging undertaking. In particular, for a marketing stimulus to be judged effective, it should lead a customer to perform an intended action (e.g., purchase a product, download an app, sign-up for a newsletter, etc.). Data on individual customers, ad exposure, and customer conversion is often available, especially in online marketing. However, a co-occurrence of customer behavior and ad exposure is insufficient to conclude that the ad caused the observed customer action. Establishing such causal link represents a major obstacle in measuring marketing effectiveness (Rzepakowski & Jaroszewicz,

2012a).

A large body of literature examines data-driven models for customer targeting in offline settings, e.g. catalog marketing, and e-commerce, e.g. real-time couponing. Literature surveys in customer relationship management (CRM) (Ngai et al., 2009) and specific CRM tasks such as churn modeling (Verbeke et al., 2012) or direct marketing (Bose & Xi, 2009) illustrate the popularity of supervised machine learning methods to develop targeting models. Using data from a past campaign including explanatory variables (e.g., customer characteristics) and a response variable (e.g., whether a customer has churned or bought an item from a sales catalog), a learning method estimates a functional relationship between the response and explanatory variables. The estimated model facilitates predicting the value of the response from the explanatory variables (e.g., for novel customers). The uplift modeling community calls this approach response modeling because the model learns to recognize customers that have responded in the past (Radcliffe & Surry, 1999). Although widely used in the literature, the response modeling approach is flawed in that it disregards causality. For example, a customer may receive a special offer and buy the advertised product subsequently, but she may have bought the same product without the discount (Radcliffe, 2007). Uplift models overcome this inadequacy through predicting differential response; that is whether the customer buys because of the offer (Kane et al., 2014). Therefore, the uplift concept quantifies the true effectiveness of a campaign (Lo, 2002).

Uplift models support marketing managers in campaign planning and targeting marketing communication to customers who would not convert without the incentive (Sołtys et al., 2015). This implies that an uplift model aims at estimating a causal link between a marketing action (e.g., offering a customer a special deal) and customer behavior (e.g., accepting the offer). Estimating the change in customer behavior that results from a solicitation, uplift models are especially suitable to support targeting decisions in campaign planning and increase campaign profitability (Radcliffe & Surry, 2011; Siegel, 2011). An analysis of the literature on uplift models in marketing reveals that existing approaches focus on conversion and churn modeling, the goal of which is to win novel customers and prevent customer defection, respectively (Park & Park, 2016; Verbeke et al., 2012). In this regard, the strategic marketing objective behind current models is market share. In terms of the underlying uplift modeling methodology, conversion and retention models predict a dichotomous response variable using classification methods.

The paper extends previous literature through introducing revenue uplift modeling. A revenue uplift model predicts the incremental revenue that results from targeting a customer with a marketing message. In many applications, customers differ in their spending (Bahnsen et al., 2015). Modeling revenue uplift accounts for this type of heterogeneity, which a conversion model is unable to accommodate. Therefore, revenue uplift modeling reflects the value-based idea of CRM (Reinartz & Kumar, 2003). Considering the focus of prior work on conversion uplift for customer acquisition and retention, revenue uplift modeling is also a relevant addition in that it provides an approach to target marketing campaigns that aim at increasing customer spending such as cross-/up-selling campaigns (Netessine et al., 2006).

In summary, the paper makes three contributions. First, existing approaches toward uplift modeling are categorized to sketch the field and highlight conceptual differences. This is useful since uplift modeling is still a niche topic in the academic literature. Second, a novel modeling strategy is proposed to predict revenue uplift. Targeting marketing communication so as to maximize revenue uplift is especially suitable for campaigns that aim at growing existing customers (e.g., cross-/up-selling). In that sense, the new approach naturally complements existing solutions for conversion and retention uplift modeling, which are geared toward customer acquisition and preventing customer defection, respectively. Third, a comprehensive empirical evaluation is carried out to demonstrate the effectiveness of the new uplift model in a real-world e-commerce environment. In addition to assessing alternative uplift modeling strategies, the experiment also provides original insights into the comparative performance of alternative machine learning methods for classification and regression to implement uplift models.

The results of the experiment confirm the effectiveness of the proposed approach. For the large e-commerce data set employed in the study, which comprises campaign results and actual sales from several e-shops across different industries, the new revenue uplift model provides the largest increase in incremental revenue and outperforms the benchmarks considered in the study. Although the model's uplift estimate is not unbiased, it's bias is somewhat lower compared to revenue models from challenger approaches because of the unique modification of the target variable. Furthermore, previous (conversion) uplift models are found ineffective in that they fail to outperform a simple response modeling approach. These results provide strong evidence that revenue uplift modeling is a useful technique to target marketing communication to responsive customers.

The paper is organized as follows: The next section introduces uplift modeling fundamentals before relevant prior work is revised. Subsequent sections elaborate on the proposed methodology and the experimental design. Afterwards, empirical results for conversion and revenue uplift modeling are reported, integrated, and discussed. The paper then concludes with a summary and outlook to future research.

4.2 Uplift Modeling Fundamentals and Process Model

The philosophy of an uplift-based targeting approach is that marketing communication should concentrate on customers who are influenced by the campaign (Rzepakowski & Jaroszewicz, 2012b). Rather than predicting customers' response probability and soliciting likely responders, as done in response/churn modeling (Chen et al., 2015; Neslin et al., 2006), the targeting decision should be based on the change in customers' likelihood to respond due to being targeted. These customers are called *Persuadables* in the literature and constitute the only group worth a marketing investment (Kane et al., 2014).

Identifying the treatment effect requires information on the response of individuals who have not received the treatment. Since each individual cannot be simultaneously treated and not-treated, the treatment effect is identified using the outcome observed in a control group. Therefore, an experimental setting with randomized treatment and control group is a prerequisite to develop

an uplift model. This may be seen as a disadvantage compared to response modeling. However, in marketing and especially online marketing obtaining control group information is relatively straightforward. In particular, A/B testing is a popular approach to perform random experiments in e-commerce. For example, a website owner may randomly assign visitors to different groups each of which get to see a different version of the homepage, hoping that the random assignment facilitates causal statements as to the effectiveness of the different page versions.

To the best of our knowledge, the literature on uplift models relies exclusively on this approach of treatment-control group comparisons to establish causality. However, it is important to note that A/B tests may fail to implement a statistically sound random experiment, especially in high-dimensional settings, which may invalidate conclusions on causal links, and may be impractical in large-scale settings where a vast number of tests are performed in parallel (Kohavi et al., 2013). For consistency with previous literature on uplift modeling, we focus on randomized trials in the form of A/B tests as vehicle to establish causal relationships. Evaluating other causal inference procedures, for example propensity scores or instrumental variables (Imbens, 2004), for uplift modeling is a fruitful area of future research but beyond the scope of this paper.

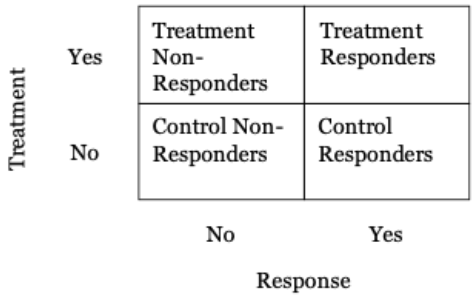


Figure 4.1: The four-fold target matrix

A/B tests are used to estimate the marginal performance increase due to a marketing incentive, the uplift, but also facilitate the training of models based on this metric (e.g. Radcliffe & Surry, 2011). In Figure 4.1, we summarize the concept of uplift with the four-fields target matrix from Kane et al. (2014). Response models distinguish between responders and non-responders (left and right column) irrespective of the actual effect of treatment. The goal of uplift models is to use the information on the control population to also account for variation in response rate dependent on whether the marketing incentive was received. In other words, uplift models identify likely treatment responders (upper right), who respond specifically due to the marketing incentive and would not respond otherwise.

To formalize the methodological difference between uplift and response modeling, let $\mathbf{X_i} = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a vector of characteristics (i.e., explanatory variables) of customer i, and let $Y_i \in \{0, 1\}$ be a binary response variable, for example whether customer i bought a product in a previous campaign. Uplift models build on the concept of A/B testing, meaning that customers are divided into two groups: treatment and control (Kohavi et al., 2009). Let $T_i \in \{0, 1\}$ be an indicator variable of the group membership of customer i, with $T_i = 0$ and $T_i = 1$ indicat-

ing membership to the control and treatment group, respectively. Then, with $P(Y_i|\mathbf{X_i}, T_i = 1)$ and $P(Y_i|\mathbf{X_i}, T_i = 0)$ denoting customer-level probabilities in the corresponding groups, traditional response models predict the conditional probability $P(Y_i|\mathbf{X_i}, T_i = 1)$, whereas an uplift model predicts the change in behavior resulting from a treatment $P(Y_i|\mathbf{X_i},\ T_i = 1) - P(Y_i|\mathbf{X_i},\ T_i = 0)$. In marketing, the treatment can be an advertisement, direct mail, or some other marketing action. Many supervised learning methods are available to estimate conditional response $P(Y_i|\mathbf{X_i})$ (Hastie et al., 2009).

An intuitive approach to develop an uplift model involves estimating two models to predict $P(Y_i|\mathbf{X_i}, T_i = 1)$ and $P(Y_i|\mathbf{X_i}, T_i = 0)$, respectively. Campaign planners can then calculate the uplift for individual customers as the difference between these models' predictions and target customers in the order of their estimated uplift. This approach is known as the two-model or indirect approach (e.g. Lo & Pachamanova, 2015). The indirect approach embodies the objective to maximize responders in the treatment group while minimizing control group responders but suffers important limitations. First, estimating two models increases computational costs. Second and more importantly, the distribution of the difference of the probabilities is often different from the distribution of the respective probabilities, which causes bias and poor model performance (Chickering & Heckerman, 2000; Rzepakowski & Jaroszewicz, 2012a).

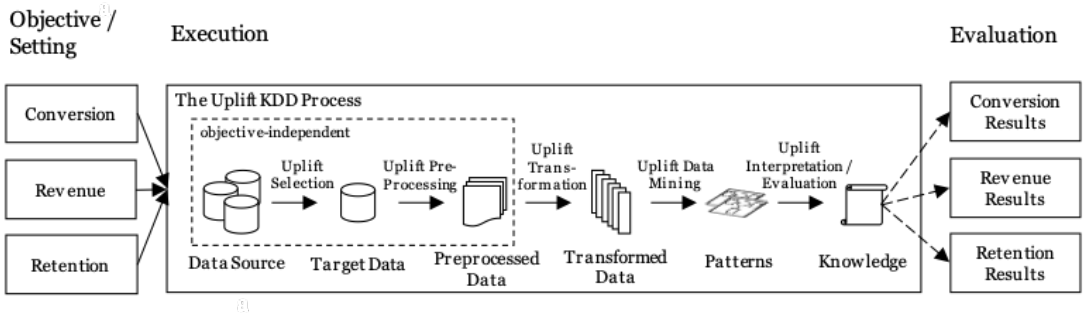


Figure 4.2: The uplift modeling process for marketing

The shortcomings of the indirect approach led to the development of improved uplift modeling regimes. Furthermore, the distinction of treatment and control group customers has implications for all stages of the model development process. To systematize related work in the field and identify the contribution of the paper, the uplift modeling process for marketing (UMPM) is introduced in Figure 4.2. The process model is based on the well-known KDD process (Fayyad et al., 1996).

Prior work on uplift modeling in marketing focuses predominantly on conversion uplift. Few studies examine retention uplift (Guelman, 2014; Siegel, 2016). Revenue uplift modeling has not been considered at all but is introduced here. The UMPM strives to raise the awareness of different modeling objectives in campaign planning. To that end, the UMPM distinguishes three stages: (1) selecting a suitable business objective out of conversion, revenue, or retention modeling for a specific campaign, (2) pursuing the chosen objective to gain insight (along the stages of the UMPM), and finally (3) evaluating results in recognition of the campaign objective to identify and target the truly responsive customers with the next marketing campaign.

All campaign goals in Figure 4.2 imply a profit objective. Winning new customers with conversion modeling increases revenues, even if the magnitude of the increase is not the focus of attention. Cross-/up-selling campaigns and campaigns aiming at customer growth in general maximize revenue directly, while campaigns to prevent customer attrition sustain future revenues. Clearly, none of the objectives and underlying uplift modeling strategies is generally preferable. Rather, the point of the UMPM is to stress that campaign planners who use uplift models to support targeting decisions should choose a definition of uplift that best matches the campaign objective and then develop a corresponding model. For example, when customer spending varies substantially and there is a small fraction of high value customers, a revenue uplift model will recommend a smaller campaign size than a conversion uplift model, which maximizes incremental sales. The smaller and more focused campaign is likely to be more profitable because it avoids the costs of soliciting low-value customers. This view is supported by the empirical results of this study.

The UMPM has been designed to provide maximal flexibility in the choice of the objective based on the respective specific marketing situations of campaign planners. Therefore, while the goal of the next campaign could be conversion-/retention-related, a rather value-related aim could be operationalized in another post-initiative (or vice versa). This single-campaign focus is typically not supported by other revenue-based models such as the customer lifetime value (CLV) which pre-empt decisions due to long-term strategies. Furthermore, CLV models are typically considered if long-term contractual agreements result from the desired action, which is rather the case for insurance or banking products/services than for fast moving consumer goods in e-retail. This is not least the case because of the accumulated value a customer generates if being locked-into a long-run agreement. In contrast, the buyer-seller relationship in e-commerce is typically rather transactional, which is why CLV models are rarely applied in this field. One might also argue that the focus on long-run customer relationships, as embodied in CLV models, is more geared toward tactic/strategic marketing management, whereas uplift models with their short-term campaign planning objectives (see Figure 4.2) are a tool for operational marketing planning. For example, measuring the causal influence of a marketing action on customer-level CLV is a complex undertaking., because changes in long-term strategic performance indicators like customer-level CLV and customer equity, respectively, can only be observed in the longer run where a multitude of external factors will simultaneously affect these indicators, leading to serious modeling issues with respect to endogeneity.

Figure 4.2 indicates that the selection of a campaign objective has methodological implications. Multiple stages in the model development process depend on the objective. Most importantly, the response variable Y_i is dichotomous in conversion and retention modeling (success/failure to convert/retain customer) and continuous in revenue modeling (purchase amount). Accordingly, conversion/retention uplift models require classification methods to estimate conditional response $P(Y_i|\mathbf{X_i})$ whereas revenue uplift models use regression methods (Hastie et al., 2009). Subsequent parts of the paper will further detail objective-specific modeling implications.

4.3 Related Literature

The review of prior work is organized along the stages of the UMPM (Figure 4.2). In general, specific modeling challenges arise in uplift modeling due to the estimation of causal effects. For example, the distinction of customers into treatment and control group affects data selection (Kane et al., 2014) as well as preparatory activities including the handling of missing values, outliers and feature selection (Hansen & Bowers, 2008; Hua, 2016; Yong, 2015). It also affects the evaluation of uplift models, which often grounds on a comparison between model predictions for treatment and control group customers (Nassif et al., 2013; Radcliffe, 2007; Radcliffe & Surry, 2011). Data transformation is important for uplift modeling because a suitable transformation of the explanatory variables or the response facilitates predicting uplift using standard learning methods (e.g. Lo, 2002; Tian & Ping, 2014).

An alternative strategy is to modify existing learning methods. In the spirit of the KDD process, an algorithmic modification exemplifies uplift data mining, which represents the prevailing approach in prior work. Corresponding studies strive to estimate uplift directly using tree-based algorithms with adapted splitting and pruning criteria (e.g. Hansotia & Rukstales, 2002), ensembles of uplift decision trees (Guelman et al., 2015), artificial neural networks (Manahan, 2005), k-nearest neighbours (Hitsch & Misra, 2018) and support vector machines (Jaroszewicz & Zaniewicz, 2016; Zaniewicz & Jaroszewicz, 2013).

The paper focuses on uplift transformation. Compared to uplift data mining, approaches for response and covariate transformation are generic. As will be detailed below, they facilitate an implementation of the modeling methodology using conventional machine learning methods. Given that this is the first paper to study revenue uplift modeling, it is useful to compare a broad set of different regression methods. Such comparison can identify methods that work well for revenue uplift. Future work could then develop modification of these methods to approach the revenue uplift modeling problem directly. In contrast, it seems less suitable to start the journey into revenue uplift modeling with a modification of one regression method, arbitrarily chosen from a vast space of alternative methods (Hastie et al., 2009).

4.4 Uplift Taxonomy

The uplift transformation framework (Figure 4.3) formally introduces and contextualizes revenue uplift modeling in the data transformation stage of the UMPM. The tree provides marketing analysts two options, a transformation of the input space (i.e., covariates) or the output space (i.e., the response variable). Response transformation can be further distinguished in terms of the underlying modeling objective.

If the objective is to increase conversion or retention rates, the response is a binary indicator variable which equals one if a customer has shown the focal behavior (has converted/churned) and zero otherwise. Response models rely exclusively on this information. Uplift models for conversion also predict a binary response variable but alter the group definition to model incremental conversions. The underlying learning methods are the same as those used in response

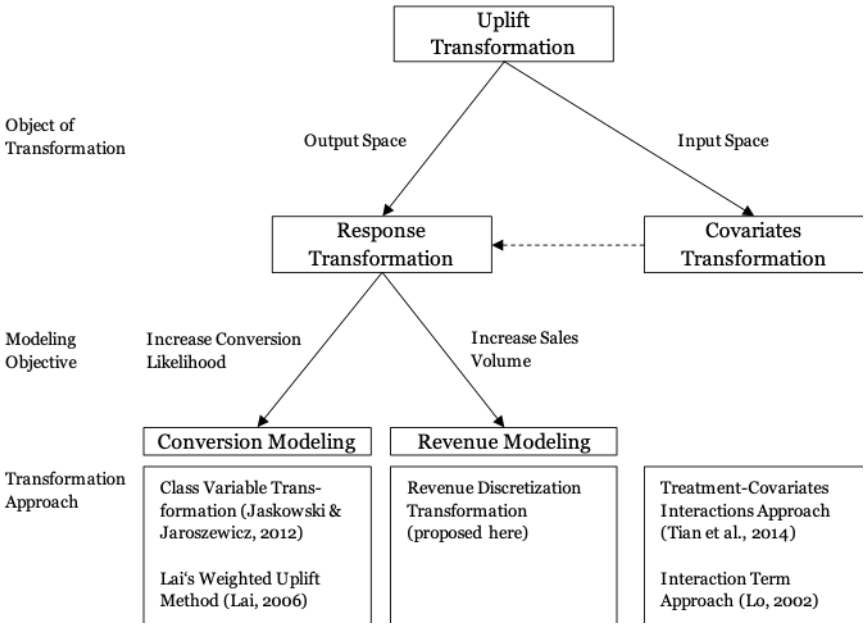


Figure 4.3: The uplift transformation framework

modeling (e.g., logistic regression, neural networks, etc.). The two main transformation approaches are the class variable transformation (CVT) (Jaśkowski & Jaroszewicz, 2012) and Lai's weighted uplift method (LWUM) (Lai et al., 2006). The paper focuses on the latter approach because recent benchmarking results indicate that it often outperforms alternative techniques (Kane et al., 2014).

Targeting models for revenue uplift transforms an originally continuous response variable (here, the revenue per customer) using information on whether customers were part of the treatment or control group. Depending on the specific transformation strategy, the new response can be continuous or binary. Drawing inspiration from previous work concerning the advantages of classification over regression models in direct marketing (Bodapati & Gupta, 2004), the proposed response discretization approach (RDT) produces a binary modeling target. However, an intermediate step in the novel approach delivers a continuous transformed response variable, which offers an alternative route to develop a revenue uplift model. A methodological difference between the two approaches is that RDT works with classification methods whereas the alternative relies on regression methods.

The literature proposes two approaches for covariate transformation; the interaction term method (ITM) (Lo, 2002) and the treatment-covariates interactions approach (TCIA) (Tian & Ping, 2014). Conceptually, both approaches are similar and differ only in the scaling of the response and normalization of the explanatory variables. In view of this, the empirical analysis includes the more recent TCIA approach.

Note that covariate transformation can be combined with response transformation. Thus, there are four options to build uplift models using covariate transformation. Either models are built on the untransformed conversion variable (conversion response modeling with modified covariates), the untransformed revenue variable (revenue response modeling with modified covariates), the

transformed conversion variable (conversion uplift modeling with modified covariates) or the transformed revenue variable (revenue uplift modeling with modified covariates). The two latter options are illustrated with the dotted arrow between the covariates transformation and response transformation boxes in Figure 4.3.

4.4.1 Conversion Response Transformation

The LWUM approach transforms the response variable so as to facilitate the use of conventional classification models to predict conversion uplift. Let $z_{i,c}$ be the binary transformed response of customer i, with c identifying the campaign objective (i.e., conversion). The response $z_{i,c}$ equals one for treatment group customers who convert and control group customers who do not convert. Both states represent a success (Lai et al., 2006). In all other cases, $z_{i,c}$ is set to zero. Formally, this logic is captured in:

$$z_{i,c} = \begin{cases} 1 & \text{if } T_i = 1 \ \cap Y_{i,c} = 1 \cup T_i = 0 \ \cap Y_{i,c} = 0 \\ 0 & \text{otherwise} \end{cases}$$

with the transformed response $z_{i,c} \in \{0,\ 1\}$. Recall that $T_i \in \{0,\ 1\}$ is an indicator variable for control/treatment group, $Y_{i,c} \in \{0,\ 1\}$ the original response variable, which captures the status of customer i (no conversion/conversion), and $i = 1, \ldots, N$ indexed customers in a past campaign of size N and $\mathbf{X_i} = (X_1, \ldots, X_n) \in \mathbb{R}^n$ is a vector of covariates. With LWUM, uplift is defined as:

$$\text{Uplift}_i^{\text{Lai}} = P\left(z_{i,c} = 1 | X_i\right) \cdot w_{\text{pos}} - P\left(z_{i,c} = 0 | X_i\right) \cdot w_{\text{neg}}$$

where w_{pos} and w_{neg} are weighting parameters determined by the ratio of positive or negative cases in the data, respectively. For $w_{\text{pos}} = w_{\text{neg}} = 1$, this approach reduces to the CVT introduced by Jaśkowski and Jaroszewicz (2012).

4.4.2 Revenue Response Transformation

From an analytical point of view, the key feature that distinguishes conversion and revenue uplift is the target variable: Instead of transforming the (binary) conversion variable $Y_{i,c}$, the (continuous) revenue variable $Y_{i,r}$ is subject to transformation. In particular, let $Y_{i,r} \in \mathbb{R}$ be the original response revenue variable capturing sales revenue of customer i, with r once again indicating the primary objective of a campaign.

The proposed RDT approach for revenue uplift modeling is based on the concept to discretize a continuous response in order to decrease the bias due to incorrect model specification and increase prediction accuracy (Bodapati & Gupta, 2004). Although the authors consider a response modeling setting, their finding appears relevant for uplift modeling as well. When correctness of a model's specification cannot be ensured, which is often the case in real-world data due to factors such as omitted variables, the resulting bias in OLS estimation can be reduced through a discretization of the target variable at the expense of an increase in variance. In large sample sizes, the importance of variance, however, diminishes. Bodapati and Gupta

(2004) gain this insight in simulation experiments with a maximum of 20,000 observations. Much larger sample sizes occur when targeting marketing communication in online environments and/or running campaigns to increase sales in e-commerce.

The logic behind value discretization can be illustrated with the sales situation of a book club (Bodapati & Gupta, 2004). Instead of predicting the annual number of books for all customers individually, the managerial challenge is to predict whether this number exceeds a pre-defined threshold. The actual task is then to determine the value of a discretizing function, $d(y)$, which the authors define as:

$$d(y) = \begin{cases} 0 & \text{if } y \in (0, y_{\text{threshold}}] \\ 1 & \text{if } y \in (y_{\text{threshold}}, \infty) \end{cases}$$

with $y_{\text{threshold}}$ as the set value of the absolute number of books in the example. Supervised classification models facilitate estimation of this function (Bodapati & Gupta, 2004).

The RDT approach proposed in this paper combines the idea of revenue uplift modeling with the target design from conversion modeling in a multi-layer transformation scheme. The revenue variable $Y_{i,r}$is first transformed[1] to obtain $z_{i,r}$ and then this variable is discretized to receive $z_{i,rg}$. More formally, the two-step transformation corresponds to:

$$z_{i,r} = \begin{cases} +Y_{i,r} & \text{if } T_i = 1 \ \cap Y_{i,r} > 0 \cup T_i = 1 \ \cap Y_{i,c} = 1 \\ -Y_{i,r} & \text{if } T_i = 0 \ \cap Y_{i,r} > 0 \cup T_i = 0 \ \cap Y_{i,c} = 1 \\ 0 & \text{otherwise.} \end{cases}$$

with $z_{i,r} \in \mathbb{R}$ as the transformed revenue that captures additional information from the group membership indicator. In particular, $z_{i,r}$ is equal to the original sales revenue for treatment group customers who made a purchase, equal to the negative sales revenue for control group customers who made a purchase, and zero otherwise. This transformation produces a novel response variable for direct uplift modeling. A single regression model suffices to predict $z_{i,r}$ which itself possesses all necessary information for uplift predictions. For RDT, however, $z_{i,r}$ is only an intermediate step. Rather than predicting $z_{i,r}$ with regression methods, a discretization procedure on $z_{i,r}$ facilitates use of classification methods and, more importantly, has the option to capitalize on the advantages of value discretization (Bodapati & Gupta, 2004):

$$z_{i,rg} = \begin{cases} 0 & \text{if } z_r \in (-\infty, 0] \\ 1 & \text{if } z_r \in (0, \infty) \end{cases}$$

where $z_{i,rg} \in \{0, 1\}$. The key differentiating factors to the discretization proposed by Bodapati and Gupta (2004) are that the response variable has been pre-transformed and that negative numbers are captured in $z_{i,rg}$, because customers who converted without having received a certain treatment are included. This points out that in $z_{i,rg}$ information related to the treatment

[1]We thank Szymon Jaroszewicz for his suggestion.

and control group is provided which underlines its characteristic of reflecting change in behavior because of having received a treatment.

The reason why the threshold has been set to zero is related to the objective in the context of uplift modeling. A "failure" is defined by $z_{i,rg} = 0$. Customers who display the behavior intended by the marketer but without having received the treatment ($z_{i,r} = -Y_{i,r}$) and customers from both treatment and control group with zero purchases ($z_{i,r} = 0$) belong to this category. In contrast, "success" is related to customers who have purchased a product with the causal connection to the campaign treatment ($z_{i,r} = +Y_{i,r}$). This group is the only one that fulfills the condition $0 < z_{i,r} < \infty$ since the price of a product always starts at one cent and is never infinite. Compared to other approaches such as CVT and LWUM, RDT defines "success" differently in terms of the four-fields target matrix presented in Figure 4.1. In this regard, the only group to target depicts the treatment responders and not, in addition, control non-responders.

4.4.3 Covariate Transformation

Covariates transformation deals with the transformation of the input space. In case of TCIA, a dummy variable $T_i^* \in \{-1; +1\}$ is created. Its value depends on whether the customer has been in the treatment or control group. Then, T_i^* is multiplied with each of the n covariates to determine the interaction term, i.e. $T_i^* \cdot X_i^*$ where X_i^* is mean centered. This additional term is taken into account when building uplift models. Following the idea by Lo (2002), the general design to model uplift is $E(Y_i|X_i) = f(T_i, X_i, T_i * X_i)$ which can be further substituted into $E(Y_i|X_i, T_i = 1) - E(Y_i|X_i, T_i = 0)$. In the next step, TCIA takes each element from the input space and transforms it using $Z_i = T_i^* X_i^*/2$ which is used to predict the response based upon the modified covariates Z_i.

4.5 Experimental Design

4.5.1 Data and Experimental Setting

The experimental setting is based on a real-time targeting process in e-commerce. When customers visit the website of an e-commerce shop, a subset of selected customers receive an e-coupon at some point during their session with a discount of 10% off the final basket value. Each targeted customer receives a unique coupon code which needs to be used during the check-out process in the basket to activate the discount. Coupons are commonly used in digital marketing to simulate conversion and generate additional sales (e.g. Khajehzadeh et al., 2014).

While clicking through the website, a visitor is randomly assigned to either being scored by a random process or by a model, i.e. all customers are subject to pre-screening determining if they are eligible for the coupon campaign. Only those customers are further considered who have a high likelihood of responding to the coupon. In the next step, customers having a high likelihood of responding are randomly assigned to the treatment or control group. Customers in the treatment group receive a coupon, those in the control group do not receive a coupon. This process provides the treatment and control setting required for uplift modelling (Figure

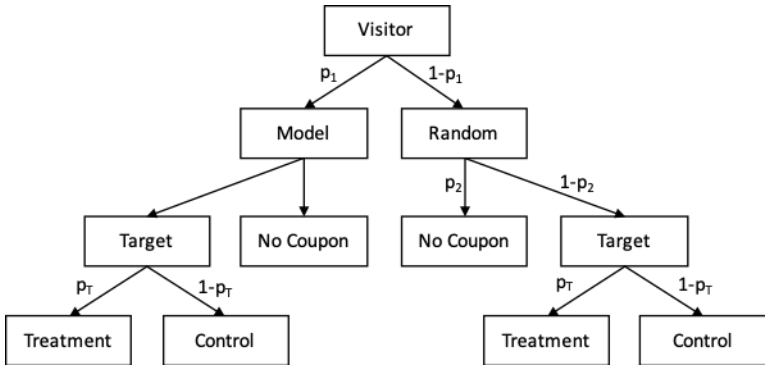


Figure 4.4: Treatment/control group assignment process for the dataset

4.4). The filtering stage, which identifies customers with a high likelihood of response, creates a selection bias towards more likely purchasers in the overall group resulting in a quasi-experiment.

A partner from industry provided the real-world data which is based on twenty-five different e- shops. There are 3,051,990 observations per variable and 62 variables. Each observation represents an individual customer session. The variables mainly capture customer-specific information such as key areas of the websites visited, including related length of time information. The data also covers the group membership indicator, shop-ID, time stamp together with information on (raw) conversions and basket values.

Table 4.1 summarizes (i) the fraction of visitors in the treatment and control group, (ii) how many visitors of each group have made a purchase, and (iii) the overall uplift on the dataset based on the group differences in conversion rates. From the table's last column, it can be concluded that the overall uplift on the real-world dataset for the experiments is low. This suggests that the specific coupon offer is not particularly effective in increasing conversion behavior. However, this does not affect the suitability of the data since the focus of the paper is on uplift modeling strategies and thus the relative gain in conversions/revenues due to an improved targeting strategy.

Since the primary objective of the paper is to introduce revenue uplift modeling and the novel RDT approach in particular, it is interesting to examine the consequences of the steps in RDT on the revenue distribution. This analysis is shown in Figure 4.5. Besides the group membership distribution (left panel), the distribution of the baseline revenue response Y_r is presented (upper right). Moreover, the two smaller plots on the bottom highlight the distributions of the transformed revenue response without discretization (z_r) and after discretization (z_{rg}), respectively. Note that the analysis is based on a sub-sample of the whole dataset (approx. 420,000 observations) which is representative to the data used in the empirical study of this paper.

The horizontal axes for both the Y_r and z_r plots extend to show the most extreme observed values, even if their frequency is too low to be displayed without scaling. Thus, for Y_r the minimum for revenue is €0 and its maximum is around €3,000 and for z_r the minimal value is -€3,000 and the maximal value €3,000 (due to the transformation logic of the first step of RDT). It is noteworthy that there are few cases where a treatment led to a purchase of a high-

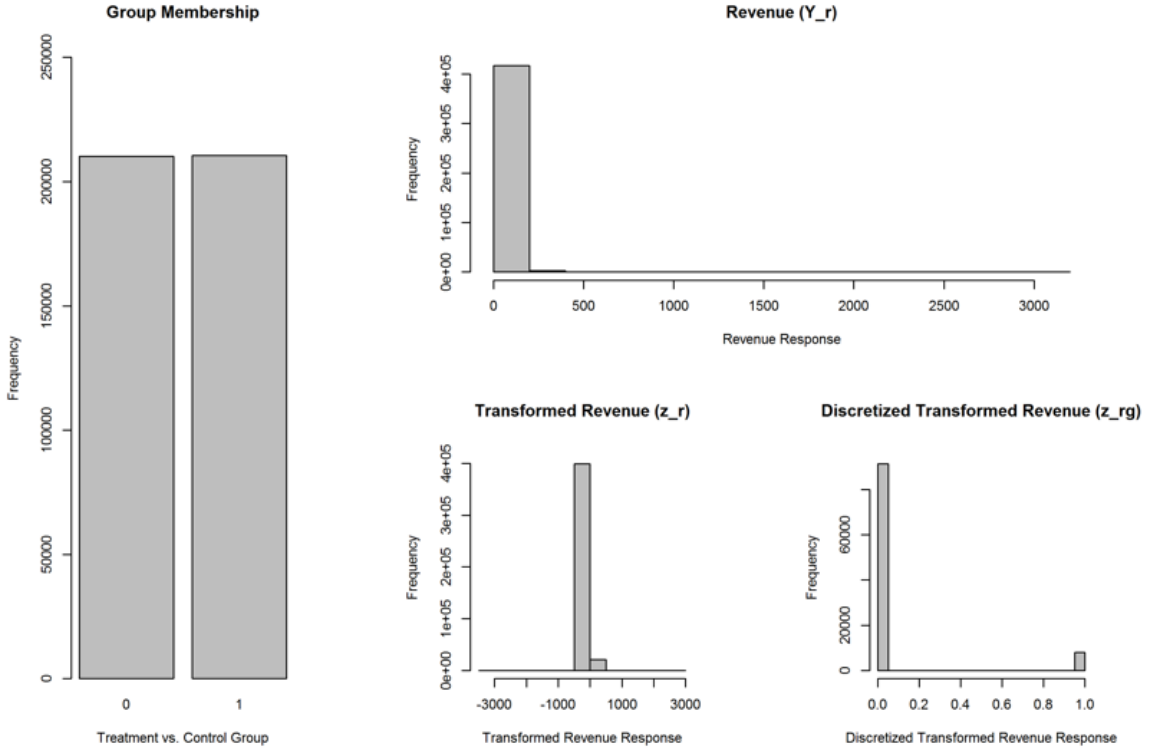


Figure 4.5: Distribution of the revenue-transformed response

Table 4.1: Average treatment effect/uplift for the dataset

Group	Share of Data	No. of Observations	No. of Converters	Conversion Rate	Uplift
Treatment	74.9%	2,285,835	175,791	7.69%	0.22%
Control	25.1%	766,155	57,285	7.47%	
Total	100%	3,051,990	233,076		

priced product. The same is true for control group customers who have purchased a product starting in the price category of €1,000 (i.e., generating high revenue hereby). This is because the high values (with and without inverted sign) occur so rarely. Since the comparably most frequent values of z_r are either negative or zero, the discretized transformed response z_{rg} is mainly zero. Only in few cases, i.e. when $z_r \in (0, \infty)$, it holds that $z_{\text{rg}} = 1$. This is visualized in the bottom right chart.

4.5.2 Base Learners

Alternative uplift modeling approaches are implemented using supervised learning methods. Table 4.2 lists the methods that have been considered in the experiments. The selection of methods includes well-established individual learners (e.g. logistic regression and tree-based learners) and ensemble algorithms (e.g. random forest and gradient boosting). Interested readers find a comprehensive description of these methods in Hastie et al. (2009). In addition, Table 4.2 includes some methods that have recently shown promising results, especially in medical and biological informatics research (Geurts et al., 2006; Soltaninejad et al., 2017, e.g. Extremely Randomized Trees) or seem to be often overlooked despite their advantages (e.g.

Theil-Sen regression, see Fernandes & G. Leblanc, 2005).

Many learning methods exhibit meta-parameters to adapt an algorithm to a particular data set (Hastie et al., 2009). Such parameters are tuned using grid-search, for which candidate parameter values have been obtained from literature (Lessmann et al., 2019).

4.5.3 Validation Strategy

The whole dataset has been partitioned into a training set (40%), a meta-parameter optimization set (30%) and a validation set (30%). In a first step, the models of all approaches are built on the training set and tested on the parameter optimization set to identify the optimal parameter configuration for the respective models. In a second step, the best models are trained on the training and parameter optimization set together (covering 70% of the whole dataset) and tested on the validation sample.

4.5.4 Performance Measures

Measures to assess predictive models are based on a comparison of actual and predicted outcomes for every individual unit of observation (Hastie et al., 2009). In uplift modeling, however, such comparison is impossible since no customer can receive and not receive a treatment at the same time (Radcliffe, 2007). This phenomenon is known as the fundamental problem of causal inference (Holland, 1986). To evaluate uplift models, Qini curves and the corresponding Qini values have been developed. They can be considered an extension of cumulative gain charts and the corresponding Gini coefficient, which facilitate an assessment of response models (Radcliffe, 2007). Gain analysis assesses models in terms of cumulative increase of responses that follow from a model-based compared to a random targeting. For the standard lift metric, *gain* is defined as the number of conversions or the value of these conversions for response and revenue models, respectively, while the uplift metric considers the *incremental* or relative gain as compared to the control group.

The performance of uplift models is visualized using Qini curves by plotting the incremental gain against the percentage of the population that is targeted. Incremental gain is determined by, first, ordering the population by their model score and segmenting customers into groups with decreasing predicted response probability. Second, the incremental gain within each segment is calculated as the difference between responders (or revenue) in the treatment group and control group adjusted for the size of the groups.

The Qini coefficient provides a single number of model performance, which is useful to compare alternative models. To calculate the Qini coefficient, the Qini curve of a model is compared to a random model (Radcliffe & Surry, 2011). The performance line of the latter starts in the coordinate system's origin and ends up in (N, n) with N as the population size and n as the total incremental number of purchases (conversion modeling) or total incremental revenue (revenue modeling) if everyone is targeted instead of a certain subpopulation (Radcliffe, 2007). The random model poses a useful baseline that relevant models need to outperform to generate value. The Qini values Q is defined as the area between the model gain curve and the random

Table 4.2: Base models

Conversion Models	Revenue Models
Logistic Regression (LogR)	Linear Regression (LinR)
Calibr. Linear Support Vector Machine (SVM)	Ridge Regression (Ridge)
k-Nearest-Neighbors (KNN)	Lasso Lars Regression (LL)
Naïve Bayes (NB)	Stochastic Gradient Descent Regression (SGDR)
Stochastic Gradient Descent Classification (SGDC)	Theil-Sen Regression (TS)
Random Forest for Classification (RFC)	Random Forest for Regression (RFR)
Calibr. Random Forest for Classification (RFC-C)	Extremely Randomized Trees (ERT)
Extremely Randomized Trees (ERT)	
Gradient Boosting for Classification (GBC)	

model (diagonal line). It can be understood as an absolute measure of incremental gain. For clarity, we denote the Qini values for the two modeling objectives, i.e. incremental number of purchases for conversion modeling and incremental revenue for revenue modeling, by Q_c and Q_r respectively.

A limitation of Q may be seen in the fact that different parts of gain/Qini curve carry different relevance to marketing practice. Campaigns are typically target to a small fraction of customers. Thus, the gain of a model for smaller targeting fractions is particularly important. Ling and Li (1998) proposed a weighting procedure to account for this issue in response modeling. We adopt their approach for uplift modeling. In formal terms, let Q_{wc} and Q_{wr} be the weighted scores across deciles of a certain model for conversion and revenue modeling, respectively.

Then, $Q_{\text{wc}} = \frac{(0.9*Q_{1,c}+0.8*Q_{2,c}+...+0.1*Q_{9,c})}{\sum_i Q_{i,c}}$ and $Q_{\text{wr}} = \frac{(0.9*Q_{1,r}+0.8*Q_{2,r}+...+0.1*Q_{9,r})}{\sum_i Q_{i,r}}$ with c and r indicating conversion and revenue, respectively, and $i = (0,\ 1,\ \ldots,\ 9)$ representing a decile index.

The following chapters present the experiments using the above performance measures. These are (i) Qini curves and Qini values Q_c and Q_r, (ii) their weighted versions Q_{wc} and Q_{wr} and (iii) incremental revenue.

4.6 Conversion Modeling

In terms of conversion modeling, we consider LWUM, TCIA and response modeling. Response modeling serves as benchmark that disregards the uplift philosophy. In total, 316 different classification models have been developed per approach, using the learning methods outlined above. Accordingly, a total of 948 classifiers are compared in the experiment. Some models have returned comparably biased probabilities. To address this problem, probability calibration based on Platt Scaling (Platt, 1999) and isotonic regression have been used for certain linear support vector machines and random forests, respectively. Figure 4.6 depicts model performance in terms of Qini curves per uplift modeling approach. The legend in each plot also provides Qini values.

Figure 4.6 indicates that most of the uplift models succeed in outperforming the naïve benchmark, which is represented by the diagonal line. However, uplift models built on top of a

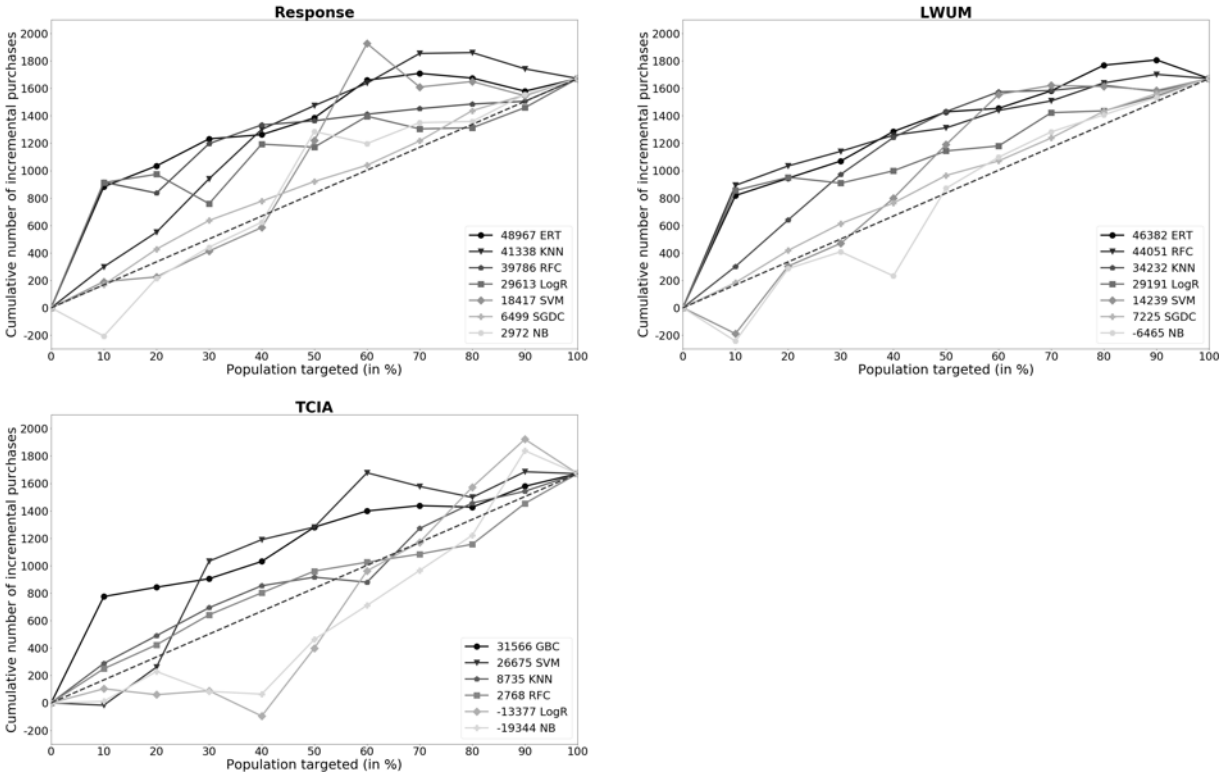


Figure 4.6: Best base models per approach for conversion modeling

naïve Bayes classifier deviate from this pattern and typically perform weaker than the naïve benchmark. In this sense, Figure 4.6 warrants the conclusion that naïve Bayes is not a suitable approach for this type of learning problem and should be avoided. Although its performance is better than that of naïve Bayes, stochastic gradient descent for classification appears to be another candidate learner which proofs inadequate for the focal prediction task. The corresponding Qini curve falls sometimes below the naïve benchmark and never exceeds it with substantial margin. On the other hand, tree-based ensemble classifiers are among the best classifiers and show consistently good results across all uplift modeling approaches. The same applies to the KNN classifier, which is always among the top three methods per approach. This result is surprising in that KNN is a rather simple classifier.

A positive result shown in Figure 4.6 is that several of the considered uplift models display a steep increase in performance within the first decile. It is common practice in marketing to target only a small subset of the customer base with a campaign. Therefore, the degree to which a model delivers high uplift in the first decile (i.e., succeeds in identifying a small subset of highly responsive customers) is of paramount importance for campaign planning practice.

To simplify comparisons of alternative uplift modeling approaches to each other, Table 4.3 reports the per-decile-uplift for each approach and classifier. In addition, the second to last and last columns provide the weighted average for conversion Q_{wc} and the rank of an approach-classifier combination across all candidates in Table 4.3, respectively. Table 4.3 reveals that the overall best approach in the comparison is a response model with underlying ERT classifier ($Q_{\text{wc}} = 5,598$). This is a stunning result, suggesting that none of the uplift models outperforms

Table 4.3: Uplift per decile by approach and base model for conversion modeling

Approach	Model	10%	20%	30%	40%	50%	60%	70%	80%	90%	Q_{wc}	Rank
Response	ERT	881	1034	1231	1261	1384	1656	1708	1673	1579	5598	1
	RFC	916	836	1197	1333	1363	1411	1451	1485	1505	5259	4
	LogR	913	973	759	1193	1171	1396	1304	1312	1459	4790	5
	KNN	299	550	940	1301	1474	1637	1854	1860	1742	4641	8
	SVM	190	225	413	585	1219	1926	1609	1649	1546	3339	11
	SGDC	168	427	636	777	921	1039	1216	1437	1551	3088	14
	NB	-206	214	442	623	1285	1197	1349	1358	1555	2622	17
LWUM	RFC	893	1034	1140	1258	1311	1439	1508	1639	1701	5365	2
	ERT	818	944	1070	1285	1428	1454	1582	1769	1807	5316	3
	LogR	855	951	909	998	1143	1180	1422	1436	1543	4676	7
	KNN	301	640	973	1241	1430	1574	1584	1622	1578	4510	9
	SGDC	185	422	606	763	981	1117	1263	1423	1559	3143	13
	SVM	-189	303	470	798	1189	1555	1621	1613	1584	3064	16
	NB	-243	286	408	232	871	1097	1282	1407	1533	2129	18
TCIA	GBC	775	843	905	1031	1281	1399	1438	1427	1579	4698	6
	SVM	-17	261	1033	1190	1281	1677	1578	1498	1685	3883	10
	KNN	288	490	694	854	916	880	1271	1457	1543	3286	12
	RFC	249	424	643	802	960	1025	1084	1157	1453	3087	15
	LogR	103	60	88	-96	399	962	1171	1573	1922	1587	19
	NB	11	229	84	64	463	710	965	1221	1838	1523	20

a simple response model. Although the latter ignores the critical point that only persuadable customers are worth targeting, the incremental conversion of the response model exceeds that of the uplift approaches, which are deliberately designed to maximize incremental response. In this sense, the results of Table 4.3 put the merit of conversion uplift modeling very much into perspective.

The second-best approach in the comparison is LWUM developed on top of a random forest classifier ($Q_{\text{wc}} = 5,365$), followed by another implementation of this approach using the ERT classifier ($Q_{\text{wc}} = 5,316$). The other uplift approach, TCIA, performs much worse and proofs inferior to Lai's approach. LWUM was the overall best approach in a recent uplift modeling benchmark (Kane et al., 2014). In this sense, superiority over TCIA, which we observe, is consistent with prior work. However, the performance of the response modeling approach remains the key finding from the conversion modeling experiment. Delivering the largest incremental gain in conversions across all but the ninths decile, which is barely relevant for marketing practice, response modeling can well be considered a dominant approach for the employed data. This sets a hard benchmark for the revenue uplift experiment using the same data, which is presented in the next section.

4.7 Revenue Modeling

The proposed RDT is deployed as candidate for revenue modeling. To demonstrate its merits, it has been tested against TCIA and the benchmark of response revenue modeling. Due to the nature of the RDT approach, i.e., the revenue response is a binary target variable after discretization, the models that have been presented for conversion modeling have been considered for predictions with this approach as well. Hence, next to 506 regression learners for response

Table 4.4: Uplift per decile by approach and base model for revenue modeling

Approach	Model	10%	20%	30%	40%	50%	60%	70%	80%	90%	Q_{wc}	Rank
Response	Ridge	1191	1205	1191	1141	1153	1213	1483	1683	1647	5563	4
	RFR	1108	1325	1231	1197	995	1108	1234	1374	1565	5379	5
	LinR	1251	962	1074	934	1339	1239	1430	1519	1615	5267	6
RDT	ERT	1470	1199	1505	1638	1396	1668	1907	1892	1717	6806	1
	RFC-C	1463	998	1361	1347	1261	1491	1592	1654	1687	6081	2
	LogR	1305	1170	874	1421	1388	1363	1397	1351	1522	5656	3
	SGDC	104	429	624	724	990	1130	1246	1414	1585	3069	7
TCIA	RFR	491	456	311	495	644	737	903	984	1274	2533	8
	Ridge	269	69	380	512	586	571	564	607	555	1738	9
	LinR	14	77	-32	-65	282	403	496	813	1087	735	10

modeling and TCIA each, additional 316 classifiers have been considered on the RDT approach, making a sum of 1,328 models for the revenue experiment.

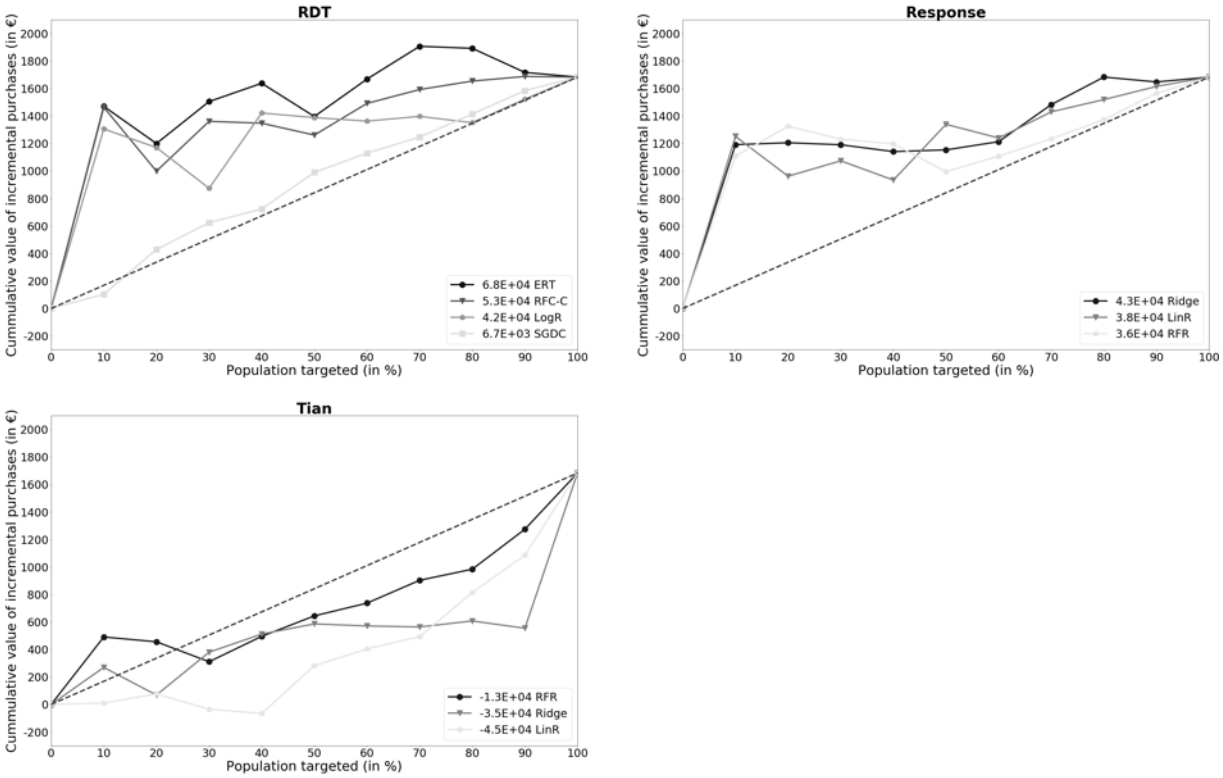


Figure 4.7: Best base models per approach for revenue modeling

This section compares the performance of stated revenue models using the described performance measures. As before from the huge model library, only those models are considered that have passed parameter optimization with greatest success, i.e. each base learner's best model. Although the subsequent Qini curves visualize the per-decile performance of the underlying models as in conversion modeling, recall that the Qini value Q_r differs to Q_c in that it identifies the value instead of the number of incremental purchases. Figure 4.7 illustrates this value as a function of the respective population's fraction for the (i) revenue response transformation (top-left), (ii) response benchmark approach (top-right) and (iii) covariates transformation for revenue modeling (bottom-left). The legends display the model values of Q_r.

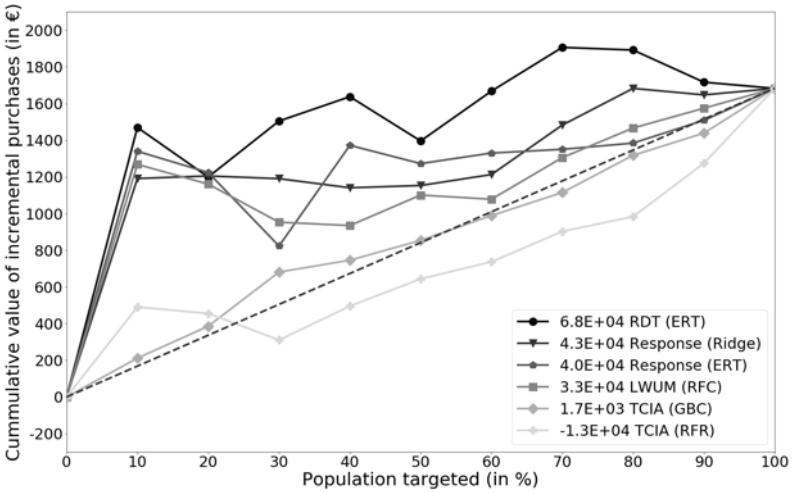


Figure 4.8: Top conversion and revenue models for incremental revenue

The performance of the shown models is summarized in Table 4.4 which reflects the results of the best models per decile and ranks them according to their weighted average for revenue. As before, for a theoretical fixed budget setting, the best approach and model combination is emphasized per decile.

Figure 4.7 and Table 4.4 clarify that the RDT approach outperforms response modeling and uplift covariates transformation on almost all deciles. At the whole, the best model on this approach, extremely randomized trees, ranks highest with $Q_{\text{wr}} = 6,806$. Remarkably, when further comparing the performance of this model with all other models for each decile separately, it outperforms on eight out of nine deciles in total. Another interesting observation is that in terms of Q_{wr}, all models of RDT rank better than all response models which, in turn, dominate all transformation-related models (with one exception pointing out to SGDC). This further enhances the reliability of our claim that the proposed approach maximizes value not just occasionally in terms of a single model. TCIA not just performs worse compared to the other approaches; for the majority of deciles it not even complies with random targeting.

4.8 Comparison Conversion vs. Revenue Modeling

While the best models have been empirically examined for the conversion and revenue objective separately, the key question now refers to whether revenue uplift modeling provides more value compared to conversion uplift modeling and response modeling. This comparison is carried out in this section to not just demonstrate the superiority of revenue modeling for this type of marketing application and campaign, but to also prove the effectiveness of the proposed approach based on incremental revenue; a performance indicator being widely used in industry. Figure 4.8 and Table 4.5 present performances of the identified best model per conversion/revenue approach from the previous analyses.

From Figure 4.8 and Table 4.5 we learn that the extremely randomized trees learner on RDT is overall superior. Across all deciles, this model clearly outperforms (1) response modeling, (2) conversion uplift modeling (i.e., random forest classifier on LWUM and gradient boosting on

TCIA) and (3) revenue uplift modeling in the shape of the random forest regressor that predicts with a transformed input space.

It is striking that of the models selected for this analysis, most of the top performers are tree-based and that, among them, ERT seems to be most valuable. Analyzing the best performance for each decile across objectives, the respective ERT models deliver the highest comparable value of incremental purchases on all deciles. The best model per decile is highlighted in bold font in Table 4.5.

Another argument in favor of the dominance of the RDT approach compared to the others stated stems from literature. Guelman (2014) suggests targeting the top ten percent most likely customers to respond positively to the campaign's treatment, i.e. only the customers from the first decile. Following this advice, RDT enhances incremental revenue comparably greatest with €1,470. This is about 10% more incremental revenue than the second-best model as stated in Table 4.5.

Table 4.5: Incremental revenue of best conversion and revenue models

Approach	Model	10%	20%	30%	40%	50%	60%	70%	80%	90%	
Response	Revenue	Ridge	1191	1205	1191	1141	1153	1213	1483	1683	1647
	Conversion	ERT	1340	1224	823	1372	1273	1330	1351	1385	1509
LWUM	Conversion	RFC	1268	1161	953	935	1101	1078	1305	1467	1575
RDT	Revenue	ERT	1470	1199	1505	1638	1396	1668	1907	1892	1717
TCIA	Revenue	RFR	491	456	311	495	644	737	903	984	1274
	Conversion	GBC	212	386	681	746	854	989	1115	1318	1439

The results confirm the contributions made in Bodapati and Gupta (2004) as discretizing revenue to apply classification models is a treasured possession we suggest campaign planners to carry in their toolboxes. According to the results of this paper, this is not just a valid but furthermore an innovative approach for extending the landscape of uplift modeling research and practice.

4.9 Conclusion

Empirical results have confirmed the proposed approach to be a valuable tool for revenue uplift modeling. For the data at hand, the parameter-optimized extremely randomized tree algorithm on RDT is most successful in identifying persuadable customers based. In other words, compared to other approaches considered in the comparison, RDT achieves the largest increase in incremental revenue. Although the model's uplift estimate is not unbiased, model building on a discretized (i.e., binary) response implies a smaller bias compared to an unmodified, continuous revenue response (Bodapati & Gupta, 2004).

More generally, the paper has reviewed several uplift modeling approaches and compared their effectiveness against each other and traditional response modeling in a large-scale experiment. Experimental results suggest that uplift modeling does not outperform response modeling in terms of conversion, whereas revenue uplift modeling does add value. Accordingly, the proposed

approach complements previous uplift modeling strategies and provides better performance when targeting marketing campaigns the primary goal of which is increasing revenue. Next to the comprehensive empirical study, the paper has developed a formalized uplift modeling process for marketing.

The applicability of the proposed model is not restricted to the online sphere. In fact, the original idea of uplift modeling stems from an offline setting. Many authors have indicated the effectiveness of uplift modeling with physical marketing incentives. These include Guelman et al. (2012, 2015) who sent out information letters and conducted outbound courtesy calls within the insurance industry, Kane et al. (2014) who point to a direct paper mail campaign and Radcliffe (2007) who uses catalogue mails in retail. If the data requirements for uplift modeling are fulfilled (i.e., random assignment of customers to the treatment group and sufficient number of samples), offline retailers such as brick-and-mortar stores can also make use of the approaches and models described here. There also exist situations where online communications take place (e.g., per e-mail), but purchases are undertaken offline (e.g. in brick-and-mortar stores), which build a bridge between online and offline interactions.

As indicated with the UMPM, revenue uplift models should be considered only if the marketing goal is to maximize incremental revenue. By comparing revenue uplift models to conversion uplift models and response models, we empirically confirmed that the former is superior if the campaign goal is revenue maximization. We may thus recommend targeting corresponding campaigns using the modeling approach proposed here. However, if customer acquisition/retention is the primary marketing goal, previous uplift approaches are probably better suited.

In future research, the discretization of the revenue response could be modified in that not a binary variable is induced but a categorical one (i.e., coarsening revenue) converting it into a multi-class classification problem. This would be especially valuable to account for broad or multimodal distributions of customer spending. Furthermore, in terms of the generalizability of RDT, it would be interesting to examine application areas other than e-couponing. A final note for future research is directed to design direct revenue uplift models that are out of the scope of transformation-based modeling architectures.

Bibliography

Bahnsen, A. C., Aouada, D., & Ottersten, B. (2015). Example-dependent cost-sensitive decision trees. *Expert Systems with Applications*, *42*(19), 6609–6619. https://doi.org/10.1016/j.eswa.2015.04.042

Bodapati, A., & Gupta, S. (2004). A direct approach to predicting discretized response in target marketing. *Journal of Marketing Research*, *41*(1), 73–85.

Bose, I., & Xi, C. (2009). Quantitative models for direct marketing: A review from systems perspective. *European Journal of Operational Research*, *195*(1), 1–16. https://doi.org/10.1016/j.ejor.2008.04.006

Chen, Z.-Y., Fan, Z.-P., & Sun, M. (2015). Behavior-aware user response modeling in social media: Learning from diverse heterogeneous data. *European Journal of Operational Research*, *241*(2), 422–434. https://doi.org/10.1016/j.ejor.2014.09.008

Chickering, M., & Heckerman, D. (2000). A Decision Theoretic Approach to Targeted Advertising, In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI'00)*, Morgan Kaufmann.

eMarketer. (2016). Worldwide Retail Ecommerce Sales Will Reach $1.915 Trillion This Year.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27–34.

Fernandes, R., & G. Leblanc, S. (2005). Parametric (modified least squares) and non-parametric (Theil–Sen) linear regressions for predicting biophysical parameters in the presence of measurement errors. *Remote Sensing of Environment*, *95*(3), 303–316. https://doi.org/10.1016/j.rse.2005.01.005

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Guelman, L. (2014). *Optimal Personalized Treatment Learning Models with Insurance Applications* (Doctoral Thesis). Universitat de Barcelona.

Guelman, L., Guillén, M., & Pérez-Marín, A. M. (2012). Random Forests for Uplift Modeling: An Insurance Customer Retention Case, In *Proceedings of the International Conference on Modeling and Simulation in Engineering, Economics and Management (MS 2012)*, Berlin, Heidelberg, Springer.

Guelman, L., Guillén, M., & Pérez-Marín, A. M. (2015). Uplift Random Forests. *Cybernetics and Systems*, *46*(3-4), 230–248. https://doi.org/10.1080/01969722.2015.1012892

Hansen, B. B., & Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, *23*(2), 219–236.

Hansotia, B., & Rukstales, B. (2002). Incremental value modeling. *Journal of Interactive Marketing*, *16*(3), 35–46. https://doi.org/10.1002/dir.10035

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The Elements of Statistical Learning* (Second). New York, Springer.

Hitsch, G. J., & Misra, S. (2018). Heterogeneous Treatment Effects and Optimal Targeting Policy Evaluation. *SSRN*.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354

Hua, S. (2016). What makes underwriting and non-underwriting clients of brokerage firms receive different recommendations? *International Journal of Finance & Banking Studies (2147-4486)*, *5*(3), 42–56. https://doi.org/10.20525/ijfbs.v5i3.278

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, *86*(1), 4–29. https://doi.org/10.1162/003465304323023651

Jaroszewicz, S., & Zaniewicz, Ł. (2016). Székely Regularization for Uplift Modeling. In S. Matwin & J. Mielniczuk (Eds.), *Challenges in Computational Statistics and Data Mining* (pp. 135–154). Switzerland, Springer International Publishing.

Jaśkowski, M., & Jaroszewicz, S. (2012). Uplift Modeling for Clinical Trial data, In *Proceedings of the International Conference on Machine Learning 2012 Workshop on Clinical Data Analysis*.

Kane, K., Lo, V. S. Y., & Zheng, J. (2014). Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. *Journal of Marketing Analytics*, *2*(4), 218–238. https://doi.org/10.1057/jma.2014.18

Khajehzadeh, S., Oppewal, H., & Tojib, D. (2014). Consumer responses to mobile coupons: The roles of shopping motivation and regulatory fit. *Journal of Business Research*, *67*(11), 2447–2455. https://doi.org/10.1016/j.jbusres.2014.02.012

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013). Online Controlled Experiments at Large Scale, In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2488217, ACM. https://doi.org/10.1145/2487575.2488217

Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery*, *18*(1), 140–181. https://doi.org/10.1007/s10618-008-0114-1

Lai, Y.-T., Wang, K., Ling, D., Shi, H., & Zhang, J. (2006). Direct Marketing When There Are Voluntary Buyers, In *Proceedings of the 6th International Conference on Data Mining (ICDM)*, Washington, DC, USA, IEEE Computer Society. https://doi.org/10.1109/icdm.2006.54

Lessmann, S., Haupt, J., Coussement, K., & De Bock, K. W. (2019). Targeting customers for profit: An ensemble learning framework to support marketing decision making. *Information Sciences*, *In Press*. https://doi.org/10.1016/j.ins.2019.05.027

Ling, C. X., & Li, C. (1998). Data Mining for Direct Marketing: Problems and Solutions, In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, Menlo Park, US, AAAI Press.

Lo, V. S. Y. (2002). The true lift model: A novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter*, *4*(2), 78–86.

Lo, V. S. Y., & Pachamanova, A. D. (2015). From predictive uplift modeling to prescriptive uplift analytics: A practical approach to treatment optimization while accounting for estimation risk. *Journal of Marketing Analytics*, *3*(2), 79–95. https://doi.org/10.1057/jma.2015.5

Manahan, C. (2005). A Proportional Hazards Approach to Campaign List Selection, In *SAS Users Group International 30 Proceedings*.

Media Dynamic, Inc. (2014). *America's Media Usage & Ad Exposure: 1945-2014* (tech. rep.).

Nassif, H., Kuusisto, F., Burnside, E. S., & Shavlik, J. W. (2013). Uplift Modeling with ROC: An SRL Case Study, In *Late Breaking Papers of the 23rd International Conference on Inductive Logic Programming (ILP'13)*.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research*, *43*(2), 204–211. https://doi.org/10.1509/jmkr.43.2.204

Netessine, S., Savin, S., & Xiao, W. (2006). Revenue management through dynamic cross selling in e-commerce retailing. *Operations Research*, *54*(5), 893–913. https://doi.org/10.1287/opre.1060.0296

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2, Part 2), 2592–2602.

Park, C. H., & Park, Y.-H. (2016). Investigating purchase conversion by uncovering online visit patterns. *Marketing Science*, *35*(6), 894–914. https://doi.org/10.1287/mksc.2016.0990

Platt, J. C. (1999). Fast Training of Support Vector Machines using Sequential Minimal Optimization. In B. Schölkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in Kernel Methods: Support Vector Learning* (pp. 185–208). Cambridge, MIT Press.

Radcliffe, N. J. (2007). Using control groups to target on predicted lift: Building and assessing uplift models. *Direct Marketing Analytics Journal*, 14–21.

Radcliffe, N. J., & Surry, P. D. (1999). Differential Response Analysis: Modeling True Response by Isolating the Effect of a Single Action, In *Proceedings of Credit Scoring and Credit Control VI*, Credit Research Centre, University of Edinburgh Management School, Edinburgh, Scotland.

Radcliffe, N. J., & Surry, P. D. (2011). *Real-World Uplift Modelling with Significance-Based Uplift Trees* (tech. rep.). Portrait Technical Report, TR-2011-1.

Reinartz, W. J., & Kumar, V. (2003). The impact of customer relationship characteristics on profitable lifetime duration. *Journal of Marketing*, *67*(1), 77–99.

Rzepakowski, P., & Jaroszewicz, S. (2012a). Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, *32*(2), 303–327. https://doi.org/10.1007/s10115-011-0434-0

Rzepakowski, P., & Jaroszewicz, S. (2012b). Uplift modeling in direct marketing. *Journal of Telecommunications and Information Technology*, *2*, 43–50.

Siegel, E. (2011). *Uplift Modeling: Predictive Analytics Can't Optimize Marketing Decisions Without It* (tech. rep.). Prediction Impact White Paper Sponsored by Pitney Bowes Software.

Siegel, E. (2016). *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die* (Revised and Updated Edition). Hoboken, Wiley.

Soltaninejad, M., Yang, G., Lambrou, T., Allinson, N., Jones, T. L., Barrick, T. R., Howe, F. A., & Ye, X. (2017). Automated brain tumour detection and segmentation using superpixel-based extremely randomized trees in FLAIR MRI. *International Journal of Computer Assisted Radiology and Surgery*, *12*(2), 183–203. https://doi.org/10.1007/s11548-016-1483-3

Sołtys, M., Jaroszewicz, S., & Rzepakowski, P. (2015). Ensemble methods for uplift modeling. *Data Mining and Knowledge Discovery*, *29*(6), 1531–1559. https://doi.org/10.1007/s10618-014-0383-9

Tian, Y., & Ping, Y. (2014). Large-scale linear nonparallel support vector machine solver. *Neural Networks*, *50*(0), 166–74. https://doi.org/10.1016/j.neunet.2013.11.014

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, *218*(1), 211–229. https://doi.org/10.1016/j.ejor.2011.09.031

Yong, F. H.-L. (2015). *Quantitative Methods for Stratified Medicine* (Doctoral Thesis). Harvard University.

Zaniewicz, L., & Jaroszewicz, S. (2013). Support Vector Machines for Uplift Modeling, In *13th International Conference on Data Mining Workshops*, IEEE.

Chapter 5

Customer Targeting under Response-Dependent Costs

Publication
Haupt, J. & Lessmann, S. (2020). Customer Targeting under Response-Dependent Costs. Working Paper. https://arxiv.org/abs/2003.06271v1

Abstract
This study provides a formal analysis of the customer targeting decision problem in settings where the cost for marketing action is stochastic and proposes a framework to efficiently estimate the decision variables for campaign profit optimization. Targeting a customer is profitable if the positive impact of the marketing treatment on the customer and the associated profit to the company is higher than the cost of the treatment. While there is a growing literature on developing causal or uplift models to identify the customers who are impacted most strongly by the marketing action, no research has investigated optimal targeting when the costs of the action are uncertain at the time of the targeting decision. Because marketing incentives are routinely conditioned on a positive response by the customer, e.g. a purchase or contract renewal, stochastic costs are ubiquitous in direct marketing and customer retention campaigns. This study makes two contributions to the literature, which are evaluated on a coupon targeting campaign in an e-commerce setting. First, the authors formally analyze the targeting decision problem under response-dependent costs. Profit-optimal targeting requires an estimate of the treatment effect on the customer and an estimate of the customer response probability under treatment. The empirical results demonstrate that the consideration of treatment cost substantially increases campaign profit when used for customer targeting in combination with the estimation of the average or customer-level treatment effect. Second, the authors propose a framework to jointly estimate the treatment effect and the response probability combining methods for causal inference with a hurdle mixture model. The proposed causal hurdle model achieves competitive campaign profit while streamlining model building. The code for the empirical analysis is available on Github.

5.1 Introduction

Data-driven prediction of customer behavior and the automation of campaign targeting are at the core of modern direct marketing (Olson & Chae, 2012). Direct marketing plays a key role in consumer markets with the continuous growth of e-commerce, at 1.8 trillion Euros globally

in 2019 (Statista, 2019) as the growth of e-commerce is accompanied by a growth in online and email advertising and in traditional print advertising, e.g. catalog marketing (Statista, 2017). To make advertising profitable, businesses have shifted away from blanket advertising and select which prospective customers to target. Targeting a customer is profitable if the positive impact of the marketing treatment on the customer and the resulting profit to the company is higher than the cost of the treatment.

Predicting the expected profit requires the estimation of the change in customer behavior if the customer is targeted, know as conditional average treatment effect (CATE) (Devriendt et al., 2018). Estimation of the CATE has been the focus of work under the label of uplift modeling (Gubela et al., 2019) and has received much attention in recent work in statistics (Athey et al., 2019; Powers et al., 2018) and machine learning (Shi et al., 2019) with the result that heterogeneous response to marketing treatment can be predicted more precisely.

However, profitable targeting must consider the effect of treating a customer in relation to the cost of treatment. Prior research tends to neglect application-specific profit and cost as decision variables and instead assume an external restriction on the number of customers to target (Ascarza, 2018; Gubela et al., 2020). While there exists work that explicitly develops targeting policies that optimize the profit of the marketing campaign (Hitsch & Misra, 2018), these policies are restricted to settings in which the cost of the treatment is known at the time of the targeting decision, e.g. the production and shipping of a catalog.

Many applications in direct marketing include costs that are uncertain at the time of the targeting decision because they are realized only when the customer accepts the marketing offer. These response-dependent costs are present whenever a marketing incentive is conditional on a profitable customer action. Companies use conditional incentives regularly in the form of discounts and the most salient applications have attracted much research, e.g. customer retention (Ascarza & Hardie, 2013; Backiel et al., 2016) or coupon targeting (Gubela et al., 2019; Sahni et al., 2016). Because the treatment cost is conditional on the customer action, the uncertainty about the customer action translates into uncertainty about the realization of the cost of the incentive. The targeting decision must then be based on comparing the expected profit to the expected cost of the marketing treatment, which is now uncertain but can be estimated. In addition to the estimation of the CATE, estimation of the expected cost requires a model of the customer decision under treatment. Despite the prevalence of targeted discounts in the industry and the focus of research on customer retention and couponing, the literature has not analyzed the targeting decision problem under response-dependent costs and lacks suitable modeling strategies to estimate both the treatment effect and customer choice efficiently.

This paper makes two contributions to the literature. First, we formally analyze the targeting decision problem under customer response-dependent costs. We show that profit-optimal targeting requires an estimate of the expected change in profit in the form of the treatment effect and estimate of the customer response probability under treatment. Second, we propose a framework to jointly estimate the treatment effect on profit and the absolute response probability. The two proposed models combine methods for causal inference with a hurdle mixture

model. We evaluate the effectiveness of our approach on a coupon targeting campaign in an e-commerce setting.

The paper is structured as follows. Section 5.2 summarizes the existing literature on the estimation of treatment effects in customer targeting and profit-based targeting policies that consider targeting costs. Section 5.3.1 formally analyzes the targeting decision under response-dependent costs. Section 5.3.2 introduces hurdle models within causal estimation frameworks as flexible models of treatment effect and customer response. Section 5.4 introduces the data and experimental design. The results of the experiment are evaluated in Section 5.5. Section 11.7 concludes.

5.2 Literature Review

Customer targeting subsumes research with the goal to identify which customers to target in order to maximize the profit of a marketing campaign. Research on customer targeting has been segmented into work on specific applications such as direct marketing and customer retention management. A starting point of our analysis is that direct marketing and customer churn are characterized by a shared decision problem, whose cost structure has implications for the design of targeting models. Direct marketing and churn management target specific customers with a marketing action through communication channels including website banners, email and print marketing, but differ in the goal of the marketing action. Direct marketing addresses customers to elicit a profitable customer response in the form of a purchase or request for a service. The existing research defines the customer response either as conversion, i.e. if the customer has completed a purchase in the period following the marketing action (e.g. Zhang et al., 2017), or as spending, i.e. how much the customer spent following the marketing action (e.g. Hruschka, 2010). Customer retention management addresses customers to avoid an unfavorable customer action and termination of the customer's relationship with the company, commonly referred to as customer churn (Mitrović et al., 2018). Positive customer action is defined either as retention, i.e. if the individual remains an active customer, or as customer lifetime value, i.e. the remaining net value of the customer to the company (Ascarza & Hardie, 2013). For our analysis, we refer to the customer action in both settings as customer *response*, which is positive in case a purchase takes place or a customer remains with the company, and to the spending or customer lifetime value as response *value*.

The fundamental decision criterion for customer targeting is the treatment effect due to the marketing action. The treatment effect is the expected change in behavior, measured on response or response value, that is caused by the marketing action. Recent studies on direct marketing and customer retention are careful to stress that the purpose of targeting is to identify the customer with the highest sensitivity to the marketing action (e.g. Ascarza, 2018; Rzepakowski & Jaroszewicz, 2012). The earlier practice to base targeting decisions on the estimate of response probability favors the targeting of natural responders rather than customers who are impacted by the marketing treatment. Hitsch and Misra (2018) show that conversion models may be profitable in practice when there exists a correlation between customers' natural propensity to

respond and their sensitivity to the marketing treatment. As there is no theoretical reason to assume such a correlation, an estimate of the response probability is generally insufficient to determine a profitable targeting policy as we clarify in the formal analysis of the targeting decision problem.

Recent research has therefore focused on the estimation of the treatment effect based on observed customer characteristics, commonly referred to as conditional average treatment effect (CATE). The general applicability of methods for treatment effect estimation has lead to developments spread across fields. An comprehensive overview over recent methodology is provided by the following studies and references therein: Devriendt et al. (2018) on *uplift* estimation in information systems, Wendling et al. (2018) for medical application, Knaus et al. (2019) and Künzel et al. (2019) for a more statistical perspective, and Athey and Wager (2017) for settings with continuous or repeated experiments.

The decision whether to target a customer in the campaign depends on the treatment effect in relation to the cost of the marketing action. While research has focused on the estimation of the treatment effect, insufficient attention has been paid to the cost structures of customer targeting. We distinguish two types of variable costs depending on the type of marketing treatment and communication channel. Applying the marketing action to a customer may entail *targeting-dependent* variable costs that arise whenever the action is taken. In practice, targeting-dependent costs arise for communication with the customer in the form of mail charges or call center fees and for the production of material treatments like catalogs (e.g. Hruschka, 2010).
An important characteristic of targeting-dependent costs is that they arise when the targeting decision is made, independent of its success. This differentiates targeting-dependent from *response-dependent* variable costs, which are incurred only if the customer responds positively after receiving the marketing treatment. Response-dependent costs arise from the design of marketing offers that are conditioned to apply only with a positive customer response. In practice, these offers take the form of free shipping on a future purchase or a discount on an existing service contract (e.g. Neslin et al., 2006). The value of the offer can be fixed, as in the case of coupon codes for free shipping, or relative to the response value, as for discounts on a monthly subscription fee. In both cases, if the customer responds negatively, for example, by terminating the existing contract, then the offer entails no cost for the company.
Beyond variable costs related to the targeting of individual customers, the implementation of a marketing campaign entails *fixed costs* for the design of the marketing action and the development of the targeting policy. While the fixed costs of the campaign are an important strategic consideration, they do not affect the operational targeting decision for individual customers.

The existence of targeting-dependent and response-dependent costs must be taken into account when designing targeting policies to maximize the profit of campaigns in direct marketing and churn. Despite the relevance of targeting costs for the targeting decision, the literature provides little discussion of customer targeting as a policy problem. Hansotia and Rukstales (2002) provide an analytical discussion of profit optimization under exclusively targeting-dependent variable costs. Targeting a customer is then profitable when the incremental value of the marketing

action is at least as high as its cost. The assumption of targeting-dependent costs is natural for print advertising and the decision rule is applied by Hitsch and Misra (2018) in the setting of catalog marketing, where the targeting cost is incurred by printing and sending a catalog. Neslin et al. (2006) formulate the campaign profit specific to customer retention campaigns including an estimate of the response value and response-dependent as well as targeting-dependent variable costs. We provide a comprehensive discussion of this formulation, its implicit assumptions and related issues and its relation to our results in Appendix 5.A and summarize our findings here.The churn campaign profit formulation includes a targeting-dependent contact cost and a response-dependent cost of the incentive to the firm in case the offer is accepted, but makes two restrictive assumptions. It implies that treatment effects are strictly positive and assumes a constant probability for customers to accept the offer when treated (Lemmens & Gupta, 2017). Assuming the same response probability for customers who receive the treatment ignores the heterogeneous sensitivity of customers to the treatment and the effect of the treatment on the expected cost, resulting in non-optimal targeting. Devriendt et al. (2019) relax the assumption of a constant response probability and discuss campaign profit from the uplift perspective, but focus on model evaluation rather than model estimation and uphold the assumption of a positive treatment effect. We add to the literature by providing a formal analysis of the general targeting decision problem, which considers variation in treatment effects over customers and guides model estimation under target-dependent and response-dependent costs.

As an alternative to a decision-theoretic approach for expected profit maximization, the literature has suggested the empirical optimization of the targeting policy (Lessmann et al., 2019). A popular approach towards empirical campaign optimization is to determine a threshold for the predicted treatment effect above which customers are targeted. Prior studies heuristically select the threshold that would have targeted the k deciles of the sample with the highest estimated CATE (Ascarza, 2018; Gubela et al., 2019; Hansotia & Rukstales, 2002; Radcliffe, 2007). The optimal proportion of the population to target can be approximated by comparing the group-wise average treatment effect for customers within each decile of the CATE estimates, since a correct ranking of customers by their CATE implies that the average treatment effect in groups with high model estimates must be higher than in groups with low model estimates. The evaluation of the model's ability to rank customers by their expected treatment effect is in line with industry practice to target a small group of the most profitable prospective customers, but ignores the cost of targeting to determine the size of the campaign. An advantage of the empirical approach is that it remains feasible when the CATE estimates are a biased or badly calibrated estimate of the ITE or when the profit and costs parameters of the campaign are unknown. When there exists heterogeneity in response value or costs, ranking the customer by their expected treatment effect ignores variation in expected profit that is not due to variation in sensitivity to the treatment. Note that response-dependent costs imply variation in expected cost even when the nominal cost of the treatment is constant. Under profit or cost heterogeneity, empirical thresholding of the treatment effect will not result in an optimal targeting policy, as we show in the empirical analysis.

In summary, we find that customer targeting in applications including direct marketing and

customer churn requires the consideration of the treatment effect and variable targeting costs. Targeting costs take the form of targeting-dependent costs and response-dependent costs, which are realized if the customer responds positively to the treatment. The next section provides an analysis of the customer targeting problem in settings that include customer-level heterogeneity in variable costs.

5.3 Methodology

5.3.1 Optimal Decision Making in Customer Targeting

The customer targeting decision problem is characterized by three components, 1) the value to the marketer conditional on the customer response, 2) the treatment cost conditional on the targeting decision and 3) the treatment cost conditional on the customer response. The existence of response-dependent costs differentiates most retention and coupon campaign settings from the cost setting discussed in previous studies (e.g. Hansotia & Rukstales, 2002), which assumes that all cost components are conditional on the targeting decision, but independent of the customer response.

Let $C_i \in 0, 1$ be a random variable indicating an action by customer i, who is described by a set of observed covariates X_i. We define $C_i = 1$ as an event with a positive impact on business profit, for example, a purchase by the customer for couponing or customer retention in churn modeling. Further, let $V_i \in \mathbb{R}^+$ be the gross profit before targeting costs that is associated with a positive customer action. V_i represents the customer lifetime value in churn prevention or the margin of a purchase in direct marketing and may show substantial variation across customers. For convenience, let $Y_i = C_i \cdot V_i$ be the observed profit of the targeting decision, excluding targeting cost. Note that $Y_i = V_i$ when $C_i = 1$ and $Y_i = 0$ otherwise. The probability of a positive response $p(C = 1|X_i)$ and the expected response value $E[V|X_i]$ are unknown at the time of the marketing decision and need to be estimated given the customer characteristics.

Recall that the variable costs split into two components, the targeting-dependent and response-dependent costs. Let c be a targeting-dependent cost that is constant and independent of the customer characteristics. Targeting-dependent costs can be contact costs, for example, mail charges. Let δ be a response-dependent cost that applies if the customer responds positively after receiving the marketing treatment. The response-dependent cost can be associated with a marketing incentive that is conditioned on a positive customer response, for example, a voucher for free shipping for the current purchase process. The expected response-dependent cost at the time of targeting depends on the probability that the customer will accept the offer. Besides, response-dependent costs may depend on the value of the response. When the marketing treatment is a relative discount, for example, in the form of 10% discount on the current purchase, the nominal discount depends on the completion of the purchase and the purchase amount. The expected offer cost then depends on the probability of a positive customer response and the value of the response. If a customer is not targeted by the campaign then no variable costs occur and $\delta = c = 0$.

Table 5.1 summarizes decision problems in target marketing by outlining their respective cost structure and anticipates the results of the decision analysis. The decision problems vary in the existence of the treatment- and response-dependent costs, the type of response-dependent incentive and assumptions about the treatment effect on response probability and value. We see that targeting-dependent costs apply to one stream of research with applications in catalog marketing (Hitsch & Misra, 2018) and online banner advertising (Diemert et al., 2018). The proposed decision framework applies under any combination of variable costs and is crucial whenever there are response-dependent costs. We further differentiate the response-dependent costs into offers with a fixed value, e.g. retention campaigns with a discount upon contract renewal (Devriendt et al., 2019), and offers with a value equal to a percentage of the response value, e.g. coupon banners in a webshop (Gubela et al., 2019). The decision analysis determines the decision variables indicated in the last column. These are the variables required to calculate the expected profit of the targeting decision in the specific setting as a result of our analysis. Note that the set of decision variables may simplify when assuming no treatment effect on the value given conversion for the first and last setting. We will discuss this assumption as a special case below.

Table 5.1: Decision problems in customer targeting and their decision variables

	Cost					
	Treat.-Depend.	Resp.-Depend.		Treat. Effect on		
Application Example		Fixed	Percentage	Decision	Value	Decision Var.
Advertisement						
Letter and Present[1]	yes	no	no	yes	no	$p(1) - p(0), R$
Online Banner[2]	no*	no	no	yes	no	$p(1) - p(0)$
Catalog[3]	yes	no	no	yes	yes	τ
Online Banner	no*	no	no	yes	yes	τ
Discount						
Print Retention Offer[4]	yes	yes	no	yes	yes	$\tau, p(1)$
Online Fixed Value	no*	yes	no	yes	yes	$\tau, p(1)$
Print Discount	yes	no	yes	yes	yes	$\tau, p(1), R(1)$
Coupon Banner[5]	no*	no	yes	yes	yes	$\tau, p(1), R(1)$
Coupon Banner	no*	no	yes	yes	no	$p(1), p(0)$

*We consider online marketing on the company's own website or in the form of email newsletters. Programmatic advertising on third party websites has a complex cost structure due to the underlying auction process.
[1]Ascarza (2018) [2]Diemert et al. (2018) [3]Hitsch and Misra (2018) [4]Devriendt et al. (2019)
[5]Gubela et al. (2019)

Consider an available marketing treatment and let T_i be a variable to indicate if the treatment was applied to customer i. We consider a single treatment and assume $T_i \in 0, 1$, where $T = 1$ indicates that the customer is targeted, the treatment condition, and $T = 0$ indicates that she is not, the control condition. The following analysis is easily extended to more than one treatment by considering multiple binary comparisons. The treatment is designed to increase the conversion probability of the customer or her value given conversion or both. Following the Neyman-Rubin potential outcome model, we indicate the potential outcomes under treatment

using $\cdot(0)$ and $\cdot(1)$. For example, $C_i(1)$ denotes the conversion outcome if customer i is targeted, whereas $C_i(0)$ denotes the conversion outcome if she is not targeted. The individual treatment effect (ITE) on profit is then $\tau_i = Y_i(1) - Y_i(0) = C_i(1)V_i(1) - C_i(0)V_i(0)$. We further distinguish between the ITE on response probability $\tau_i^C = C_i(1) - C_i(0)$ and the ITE on response value $\tau_i^V = V_i(1) - V_i(0)$.

We now begin our analysis of the targeting decision problem. The profit π_i for an individual in the marketing campaign including treatment costs is

$$\pi_i = \begin{cases} C_i(0)V_i(0) & \text{if } T_i = 0 \\ C_i(1)V_i(1) - C_i(1)\delta - c & \text{if } T_i = 1 \end{cases}$$

The general decision problem whether to target a specific customer under response-dependent costs can then be posed as

$$p_i(1)(V_i(1) - \delta) - c > p_i(0) \cdot V_i(0), \tag{5.1}$$

where we use p_i as a convenient shorthand for $p(C = 1|X = x_i)$. Note how the variable costs affect the campaign profit. The target-dependent costs c are realized before the customer makes any decision and are therefore independent of the customer action. The response-dependent costs δ are realized only when a positive response takes place.

Solving the inequality for the treatment effect yields

$$p_i(1)V_i(1) - p_i(0)V_i(0) > p_i(1)\delta + c \tag{5.2}$$

The optimal decision naturally depends on the individual treatment effect on the profit on the left side of the equation. However, it also depends on the probability of a positive customer response under treatment as a mitigating factor on the offer cost. Intuitively, the absolute offer costs are a promise from the firm and must be discounted by the chance that the promise will in fact be redeemed by the customer. If the customer does not redeem the offer, then the response-dependent costs are not incurred by the company. The customer targeting decision under response-dependent costs thus differs from the case where $\delta = 0$ because the costs are now stochastic rather than known at the point of the targeting decision.

The optimization of expected profit underlying Eq. 5.2 implies that, when faced with two customers with an identical CATE, it is more profitable to target the customer who is less likely to respond positively and accept the marketing offer. The previous practice to target customers with a high response probability after treatment not only disregards the causal effect of the treatment, as previous literature has pointed out (Ascarza, 2018), but increases the cost of campaigns by targeting customer with high expected response-dependent cost. To clarify the intuition behind this result, consider the treatment of a customer as an investment with probabilistic cost. If the payout of two investments is identical, a rational agent prefers the

investment that has lower expected cost. This result suggests that when there is little or no treatment heterogeneity, meaning that the payout of the treatment is identical between customers, it is profitable to target customers with a lower rather than higher probability to respond.

In practical terms, any decision setting with response-dependent costs will require an estimate of the treatment effect $p_i(1)V_i(1) - p_i(0)V_i(0)$ and an estimate of the response probability $p_i(1)$. This result is surprising because previous literature has emphasized uplift models, which provide an estimate of the treatment effect, as a direct replacement of response models, which provide an estimate of the conversion probability. The decision under response-dependent costs requires both a model of the treatment and a model of the conversion probability under treatment.

In application, a positive expected profit may not result in an optimal policy under strategic considerations. Actual targeting campaigns are regularly evaluated by their return on advertising spend (ROAS). The ROAS is defined as the ratio of campaign profit over campaign costs. Note that the same information is sometimes expressed by its inverse as the cost-revenue ratio. The ROAS is a metric of advertising efficiency and as such does not consider campaign size. While it is generally not profit-optimal to maximize efficiency at the cost of targeting fewer customers, a minimum ROAS is often required in practice to satisfy management goals and allocate resources efficiently between marketing channels or campaigns. A side result of our analysis is that the proposed decision rule can be used to set targeting thresholds to reflect a minimum ROAS as

$$\frac{p_i(1)V_i(1) - p_i(0)V_i(0)}{p_i(1) \cdot \delta + c} \geq \text{Target ROAS}$$

We go on to discuss two special cases that arise in digital applications.
First, assume that $c = 0$. In digital marketing settings, there are no variable contact costs if customer communication is digital and automated. In particular, the costs for email targeting and banner campaigns on company's own websites arise in the form of fixed cost into infrastructure, e.g. content management systems and content production. These costs are irrelevant for operational targeting decisions in the short run. The targeting rule is then

$$\begin{aligned} p_i(1)V_i(1) - p_i(0)V_i(0) &> p_i(1) \cdot \delta \\ p_i(1)(V_i(1) - \delta) &> p_i(0)V_i(0) \end{aligned} \tag{5.3}$$

Assume additionally that offer costs depend linearly on the response value, i.e. $\delta_i = \eta V_i(1)$. The latter assumption corresponds to discount coupons that reduce the checkout amount by a fixed percentage, e.g. 10%, and other forms of dynamic pricing. Percentage discount coupons are frequently used in online marketing as a transparent means to differentiate incentives according to the value of customers and as an incentive that encourages higher spending. The decision

rule for discount offers requires an estimate of the expected response value under treatment:

$$\begin{aligned} p_i(1)V_i(1) - p_i(0)V_i(0) &> p_i(1) \cdot \delta_i \\ p_i(1)V_i(1) - p_i(0)V_i(0) &> p_i(1) \cdot \eta \cdot V_i(1) \end{aligned} \tag{5.4}$$

Second, there exists a special case of the decision problem in Eq. 5.4 that requires no estimate of the purchase value. Assume that the treatment affects the conversion probability but not the response value, i.e. $V(1) = V(0)$. Then equation 5.4 reduces to

$$\begin{aligned} (p_i(1) - p_i(0)) \cdot V_i &> p_i(1) \cdot \eta \cdot V_i \\ p_i(1) &> \frac{p_i(0)}{1 - \eta} \end{aligned} \tag{5.5}$$

Note that under the combined assumptions of a percentage discount with no fixed contact cost and no effect on conversion value, the decision rule becomes independent of the individual purchase value. Intuitively, a negligible communication cost removes the need to make up for the cost of customer targeting. Further making the coupon cost dependent on the response value automatically adjusts the cost to decrease with smaller response values and vice versa. In practice, this setting requires estimation of the purchase probabilities with and without treatment.

The two special cases imply that the cost structure, which is determined by the infrastructure of the campaign and the design of the treatment, can increase or reduce the complexity of the decision problem. In general, when the cost of the treatment is conditioned on additional variables, then the estimation of these variables is relevant for the decision problem. We can see that percentage discounts on the purchase value introduce an estimate of the purchase value under treatment into the decision (Eq. 5.4). Similar arguments can be made for more specialized coupon design like a minimum purchase value or a staggered discount increasing with purchase value. The second case shows that specific cost structures may simplify the decision problem. Under the additional assumption of no treatment effect on value, the targeting decision reduces to the estimation of the probabilities of purchase with and without treatment in Eq. 5.5.

The proposed decision framework is a generalization of marketing decision settings discussed in the literature. Prior research in marketing has considered campaigns with treatment-related but no response-related costs, such as traditional mail catalog marketing (Hansotia & Rukstales, 2002; Hitsch & Misra, 2018). Assuming $\delta = 0$, we can show that the treatment effect on profit Y_i is sufficient for the targeting decision in these cases, which reduces to

$$p_i(1)V_i(1) - p_i(0)V_i(0) > c \tag{5.6}$$

We see immediately that an estimate of the treatment effect on the profit $p_i(1)V_i(1) - p_i(0)V_i(0) = Y(1) - Y(0)$ is a sufficient decision criterion under the conditions of Eq. 5.6. If we assume no treatment effect on the value such that $V_i(1) = V_i(0) = V_i$ and assume V_i to be known or

modeled independently, we recover

$$p_i(1) - p_i(0) > \frac{c}{V_i}, \tag{5.7}$$

where the focus lies on the estimation of the treatment effect on the customer response. This recovers the estimation problem addressed by prior research under the label of uplift modeling, although the dependency on the response value is not typically discussed in the literature (Devriendt et al., 2018).[1]

In summary, the treatment effect is not sufficient for profit-based targeting in settings with response-dependent costs. The additional decision variables required for the targeting decision depend on the cost-structure of the marketing treatment as given in Table 5.1. For variable costs with a fixed value, the purchase probability under treatment determines the cost as in Eq. 5.2. For treatment with a value relative to the purchase value, the purchase probability under treatment and the purchase value under treatment jointly determine the effective treatment cost as in Eq. 5.4. Both cost structures are common in direct marketing. The following sections discuss a model specification to estimate the cost-related decision variables $p(1)$ and $R(1)$ within the model of the treatment effect $Y(1) - Y(0)$.

5.3.2 Causal Hurdle Models

The targeting decision under response-dependent costs (Eq. 5.2) requires estimates of the treatment effect $Y(1) - Y(0)$, the response probability under treatment $p_i(1)$ and, for discount coupons, the response value under treatment $V_i(1)$. Alternatively, we can decompose the profit using $Y_i = C_i \cdot V_i$ and estimate the treatment effect as $p_i(1)V_i(1) - p_i(0)V_i(0)$. This formulation makes explicit that the additional decision variables are contained within the treatment effect on profit. The remainder of the study develops a framework to simplify the modeling task based on this observation.

A straight-forward approach to estimate the decision variables is to build several models, where one (causal) model estimates the CATE $\hat{\tau}_i$ and additional models to estimate remaining decision variables under treatment. In the following, we will call this the *distinct modeling* approach. The distinct modeling approach requires one model to estimate $p_i(1)$, a second model to estimate $V_i(1)$ in the case of discount coupons, and one to four models depending on the specific approach to estimate the CATE. In other words, the distinct modeling of each decision variable introduces up to two additional models to the CATE model.

In the following, we propose a framework to avoid additional model complexity and simultaneously estimate the treatment effect on expected profit, the purchase probability and purchase value. The proposed framework exploits the decomposition of the expected profit into the conversion probability $p_i(1)$ and purchase value $V_i(1)$ to collect estimates for $p_i(1)$ and $V_i(1)$ from the treatment effect model. We estimate $V_i(1)$ and $p_i(1)$ jointly within the profit model by modeling the observed profit from customer Y_i as a two-stage hurdle structure.

[1] See Ascarza (2018, fn. 27) for a brief mention of the issue in the case of customer churn.

Hurdle models, known as Tobit II models in econometrics, are mixture models over two distributions, one of which has a point mass at zero, which were previously applied in applications of customer choice (Donkers et al., 2006; Van Diepen et al., 2009). They are convenient to model decisions that involve a binary decision on whether to act, the hurdle, and a conditional decision on the value associated with acting. Hurdle models assume that the occurrence of zeros is entirely driven by a first-stage process, i.e. the second stage value is zero when the first stage decision is a negative response and strictly positive when the first stage decision is a positive response.

The hurdle model allows us to decompose the estimation of the profit Y into the estimation of response C and response value R. The probability mass function of the hurdle model is

$$\Pr(Y_i = y|X_i = x) = \begin{cases} (1 - \Pr(C = 1|X_i = x)) \cdot 0 & \text{if } Y_i = 0 \\ \Pr(C = 1|X_i = x) \cdot \Pr(V_i = v|X_i = x) & \text{if } Y_i > 0 \end{cases} \tag{5.8}$$

where $\Pr(C_i = 1|X = x_i))$ is a model for customer response and $\Pr(V_i = v|X_i = x)$ is a model of response value. If a customer chooses to respond, they decide on their spending behavior in the second stage, which determines the response value to the firm. The profit from a response is zero if a customer chooses not to respond and strictly positive otherwise.

The hurdle model specification has two properties that are relevant in the context of customer choice. First, the separation of the purchase decision and value decision facilitates the estimation and interpretation of each model. In the context of treatment estimation, separating the effect on response probability and response value provides a more nuanced understanding of marketing effectiveness and can be used to improve the treatment. The model structure also accommodates differences in, for example, the relevance of available covariates for each decision step (Donkers et al., 2006). Second, the models for the prediction of response probability and response value can be estimated separately when the purchase incidence is observed and we assume independent error terms (Cameron & Trivedi, 2005, p.545). This property will provide additional flexibility when estimating the proposed causal hurdle model in practice.

It remains to integrate the hurdle model into a framework for causal inference. Under the common assumptions of the potential outcome framework, i.e. unconfoundedness, overlap and stable unit treatment value, the CATE can be expressed as the difference between the outcome Y_i conditional on treatment assignment T_i and covariates X_i,

$$\tau(x) = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y|X_i = x, T = 1] - \mathbb{E}[Y|X_i = x, T = 0]. \tag{5.9}$$

We integrate the hurdle model into the standard treatment effect model by modeling profit with

the hurdle model

$$\Pr(Y_i = y|T_i = t, X_i = x) = \begin{cases} 1 - \Pr(C = 1|X_i = x, T_i = t) & \text{if } Y_i = 0 \\ \Pr(C = 1|X_i = x, T_i = t) \cdot \Pr(V_i = v|X_i = x, T_i = t) & \text{if } Y_i > 0 \end{cases} \tag{5.10}$$

where both the conversion probability and the purchase value conditional on conversion depend on the treatment assignment T_i and the covariates X_i.

Following the definition of the hurdle model above and under the assumptions of the potential outcome framework, we specify our causal hurdle model as

$$\begin{aligned} \hat{\tau}(X_i) &= \hat{y}(X_i, 1) - \hat{y}(X_i, 0) \\ &= p(C = 1|X_i, T = 1) \cdot \mathbb{E}[R|X_i, T = 1] - p(C = 1|X_i, T = 0) \cdot \mathbb{E}[R|X_i, T = 0]. \end{aligned} \tag{5.11}$$

Estimating the treatment effect on response probability and response value separately has an additional advantage if we expect heterogeneity of effect direction and size on customer value and response probability. This is the case if individual customers react differently to the same offer, for example, purchase their basket with higher probability or put additional products into their basket in response to receiving the treatment. Further, we expect the treatment effect on probability and value to be closely connected to the design of the marketing action. Under strong heterogeneity, we expect some customers to react to the marketing treatment by increasing the response value, e.g. putting more products into their basket, while becoming more reluctant to respond at the higher value, e.g. abandon a high-value shopping basket. Explicit estimates of the disentangled treatment effects are then relevant for treatment selection and design.

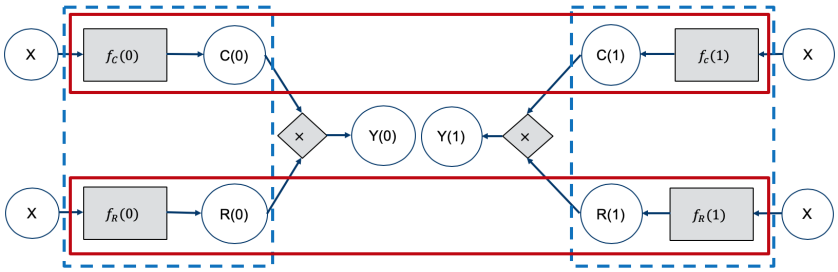


Figure 5.1: Causal hurdle model structure. Frames indicate the two proposed strategies for estimation in the form of two hurdle models (dashed blue) or two causal single models (solid red)

The formulation in Eq. 5.10 does not restrict the specific method of causal inference. Figure 5.1 visualizes the general structure of causal hurdle models and makes the estimation targets explicit. We see that one strategy to estimate all relevant decision variables is to estimate four separate models, i.e. one model each for purchase probability and purchase value times one model each for the treatment and control group. This two-model hurdle model is equivalent to combining the two-model approach for CATE estimation with two hurdle models for which the

choice and value components are estimated separately (Cameron & Trivedi, 2005, p.545).

It is possible to simplify the estimation by estimating more than one decision variable jointly. Eq. 5.9 is the starting point for two approaches to integrating a hurdle model structure into treatment effect models. Figure 5.1 visualizes the proposed methods to reduce the number of separate models by joint estimation of variables horizontally (solid red), over treatment and control group, or vertically (dashed blue), over purchase probability and value.

The *single-model hurdle model* combines the single model approach for causal inference with a two-stage estimation procedure for the hurdle model. A general model for the conditional profit with or without treatment takes the form $y = f(x, t)$ and predicts the return given the covariates X and treatment assignment T. Despite its simplicity, this *single model* approach has been found to provide competitive CATE estimates for sufficiently flexible specifications of $f(\cdot)$ (Künzel et al., 2019). The single two-stage approach estimates one model for the response probability and one model for the customer value each jointly over the control and treatment group (Figure 5.1, solid red). Following the single model approach, we include the treatment variable as a covariate into the model. By choosing a flexible parametrization $f(\cdot)$, we can model the conditional average treatment effect through the interaction between T_i and covariates X_i within the model (Hill, 2011).

5.4 Experimental Design

We evaluate the proposed methodology in an online couponing setting[2]. The decision analysis summarized in Table 5.1 identifies online couponing as a particularly interesting decision problem because showing a coupon banner to the customer entails no targeting-dependent cost while the coupon value constitutes a substantial response-dependent cost. However, studies on online couponing are scarce in the literature and we are not aware of research considering its cost setting (e.g. Gubela et al., 2019; Sahni et al., 2016).

A German online fashion retailer deploys an automated targeting system that can show website visitors of the online shop an offer for a discount on their purchase. Targeted customers receive a coupon code that provides a discount of €10 at checkout. The code is made available to the customer through a banner on the webpage, which states the discount offer and displays a coupon code to be entered during the checkout process. The banner is shown repeatedly on subsequent page views within the same session to ensure that the customer is aware of the offer. The discount is subject to common terms and conditions that require a minimum checkout value of €50 for the coupon to be usable. The operational question of the fashion retailer is to identify the customers whose incremental margin when being targeted is strictly larger than the expected cost due to the coupon. As the theoretical analysis shows, the profit-optimal targeting policy differs from similar target marketing settings discussed in previous studies as the expected cost of the coupon depends on the customer's purchase probability.

[2]The code for the experiment and evaluation is available at https://github.com/johaupt/response-dependent-costs

The data contains information on 118,622 anonymized website visitors in the form of 50 variables collected through tracking software and the shop system. Variables include information on the user history, e.g. the number of previous visits, the behavior on the website, e.g. the number of clicks on the website, and the current shopping basket, e.g. the number of items and their total price (see Baumann et al., 2019, for a comprehensive discussion of covariates). 9% of website visitors convert and complete their purchase with a median purchase value of €75. We remove 1,459 outliers with a substantially higher basket value between €300 and €1750 corresponding to the 2.5% percentile of the purchase value distribution.

The data fulfills the assumptions of the potential outcome framework. The unconfoundedness and overlap assumptions are met by design through randomizing treatment assignment, which is common practice in customer targeting applications. The stable unit treatment assumption value requires that social interaction effects between individuals are rare or small in size. This assumption has recently been challenged for the telecommunications industry (Ascarza et al., 2017). The social network that customers form when communicating via the telecommunication network have been found to lead to substantial positive spillover effects from the targeted customers to their connections. In online shopping, while there is potential for social effects, e.g. sharing information about the availability of coupons on the website, there are no social mechanisms inherent to the purchase process. We therefore assume that any potential social effects are insubstantial for our analysis, but encourage additional research on social effects in couponing applications.

Evaluation of approaches including the estimation of treatment effects is complicated by the fact that the true ITE is unobservable, so comparable studies rely on simulated data (e.g. Athey et al., 2019; Nie & Wager, 2017). To facilitate the evaluation of the proposed approach through a setting where the treatment effect is known, we conduct an empirical Monte Carlo study combining the observed covariates X and customer spending Y with a simulation of ITE (Nie & Wager, 2017). We simulate the overall treatment effect $\tau(X)$ as a combination of the treatment effect on the conversion probability $\tau_C(X)$ and the purchase value conditional on a purchase $\tau_R(X)$. Each treatment effect is determined by a linear combination of covariates with coefficients drawn randomly following

$$\begin{aligned}\beta_C, \beta_R &\sim \mathcal{N}(0_k, I_k)\\ \tau_C(X) &= X_\tau^\top \beta_C\\ \tau_R(X) &= X_\tau^\top \beta_R\end{aligned}$$

where X_τ is a subset of $k = 11$ selected variables from the full set of variables. Both treatment effects are centered and scaled. To simulate a realistic marketing setting, we scale the ITE distribution to have most of its mass in the range [0;10] and a positive average effect (e.g. Hitsch & Misra, 2018, Figure 12). For the ITE on response probability, we center the distribution around an ATE of 5 percentage points and truncate the simulated values to the range [-0.1,0.15]. For the ITE on response value, we center the distribution around an ATE of €1 and truncate

the simulated values to the range [-10,10].

We simulate the potential outcome with and without treatment by flipping the observed outcome label for observations in the treatment group chosen randomly in proportion to their $\tau_C(X_i)$ as in Nie and Wager (2017). We do not observe the potential checkout amount for 4680 customers whose outcome we flip from non-converted to converted. We choose not to remove these customers and instead approximate the data generating process of the checkout amount to generate synthetic values. We employ a gradient boosted tree ensemble (GBT) on the customers for which we observed the checkout amount to ensure that the approximating model is sufficiently flexible and predict the unobserved checkout amounts using the tree ensemble. The treatment effect τ_R is then added to the observed, or if unavailable to the synthetic, basket value. The empirical Monte Carlo approach allows us to evaluate our approach against the actual distribution of customers including the real purchase process while controlling the individual treatment effect for evaluation.

We use five-fold cross-validation to compare the causal hurdle model and the distinct modeling approach on the holdout data. Considering the estimation strategies for the treatment effect and the conversion probability results in the eight combinations summarized in Table 6.2. The proposed approaches to model the treatment effect using hurdle models are grouped as causal hurdle models. They are defined by a two-stage hurdle approach estimating the response probability and the response value given a positive response, with and without treatment. The hurdle specification provides an estimate of the conversion probability under treatment without the need to estimate an additional model. The distinct modeling approaches are defined by the estimation of the treatment effect on profit in one stage. Approaches that estimate the treatment effect on profit require the estimation of a separate model to estimate the conversion probability under treatment. For each approach, we compare a linear specification to a more flexible specification using the GBT. To simplify the analysis for the distinct modeling approaches, we choose the same specification for the models of the treatment effect and the conversion probability.

Table 5.2: Summary of model specifications considered in the experiment

	Architecture			Number
Stages	CATE Model	Conversion Model	Estimator	of Models
Causal Hurdle Models				
hurdle	single-model	-	gbt	2
hurdle	two-model	-	linear	4
hurdle	two-model	-	gbt	4
Distinct Modeling Approaches				
one-stage	single-model	separate	gbt	2
one-stage	two-model	separate	linear	3
one-stage	two-model	separate	gbt	3
one-stage	dr	separate	linear	5
one-stage	dr	separate	gbt	5

We consider three approaches for the estimation of the CATE. First, the single-model approach

that includes the treatment variable into the model. We test the single-model approach only in combination with the GBT specification because the approach requires a sufficiently flexible model to capture interaction effects between the treatment indicator and covariates. The single-model approach requires the estimation of two models. Under the hurdle model approach, the two models are one single-model including the treatment indicator respectively for the conversion and spending given conversion. Under the distinct modeling approach, the two models are a single-model for the profit and a separate model for conversion under treatment.
Second, the two-model approach that relies on the estimation of separate models for the treatment and control group. For the distinct modeling approach, the two models estimate the expected profit in the treatment and control group, respectively, with a separate model for conversion under treatment. For the hurdle approach, the two models for the treatment and control group are hurdle models that each consist of one model for the conversion and one model for the spending given conversion.
Third, the doubly-robust outcome transformation (DR) due to Robins and Rotnitzky (1995). The DR approach provides an additional benchmark that has shown strong empirical performance in the econometric literature (Knaus et al., 2019). Under the DR approach, the treatment effect is estimated using a single model on a transformation of the profit

$$Y_i^{DR} = \mu_1 - \mu_0 + \frac{T_i(Y_i - \mu_1)}{p(T=1|X_i)} - \frac{(1-T_i)(Y_i - \mu_0)}{1 - p(T=1|X_i)}$$

$$\text{with} \quad \mu_1 = E[Y|X_i = x, T_i = 1] \quad \text{and} \quad \mu_0 = E[Y|X_i = x, T_i = 0]$$

The expected profit in the treatment and control group, $E[Y|X_i, T_i = 1]$ and $E[Y|X_i, T_i = 0]$, and the probability to receive treatment, $p(T = 1|X_i)$, are estimated by three auxiliary models. For simplicity, we use linear regression to estimate the expected profit and logistic regression to estimate the probability to receive treatment.

5.5 Empirical Results

Recall that each targeting policy is a combination of an estimate of the treatment effect and an estimate of the treatment cost. The profit generated by the targeting policy depends on the quality of the estimates of the treatment effect and on the quality of the estimates of the conversion probability under treatment. The analysis is therefore structured around the evaluation of the treatment effect estimation and the evaluation of the estimation of the expected individual-specific cost in the proposed hurdle framework. First, we test if the conversion probability estimates are sufficiently informative and economically relevant for profitable targeting. We propose that the expected individual cost is practically relevant and the estimate $p(C|X_i, T = 1)$ is sufficiently precise for profitable targeting. We test their economical relevance through an evaluation of campaign profit. Second, we test if the CATE estimates in the proposed causal hurdle framework are equivalent to CATE estimates under the conventional modeling strategy. The campaign profit under joint model estimation is expected to be at least as high as under the distinct model approach, while being easier to manage in application. Therefore, we evaluate the CATE estimates using statistical metrics on the simulated treatment

effect and the evaluation of campaign profit under population-based cost estimates.
Third, we test if the proposed analytical targeting policy has a higher return than empirically optimized policies. The analytical targeting policy the individual treatment effect and response probability that requires a combination of treatment effect estimation with individual-level cost estimation. We test the campaign profit of the policy under the distinct estimation approach and the proposed causal hurdle framework.

The incremental campaign profit must be determined against baseline policies. As a general baseline, we select the sum of profit from individuals in the data when no campaign is run, i.e. no individual is targeted with the marketing treatment. We compare the proposed analytical targeting policy against the alternative *Empirical* policy suggested by our literature review, which determines a targeting threshold that maximizes campaign profit on the training data (Ling & Li, 1998).

5.5.1 Profit Implications of Individual Cost Estimates

Table 5.3: Policy profit for the conversion models evaluated under selected treatment effect estimation methods

	Architecture			Profit	Fraction Treated
Policy	CATE Model	Conversion Model	Estimator		
Baseline	-	-	-	46,236	0.00
Analytical	ATE	Conversion Rate	-	50,830	1.00
Analytical	ATE	Single-Model	GBT	52,931	0.84
Analytical	ATE	Two-Model/Distinct	Linear	51,936	0.76
Analytical	ATE	Two-Model/Distinct	GBT	52,402	0.79
Analytical	Actual	Conversion Rate	-	55,493	0.71
Analytical	Actual	Single-Model	GBT	56,696	0.72
Analytical	Actual	Two-Model/Distinct	Linear	57,361	0.69
Analytical	Actual	Two-Model/Distinct	GBT	57,022	0.69

We begin with the prediction of conversion probability under treatment to calculate the expected cost of targeting. Table 5.3 reports the profit and the fraction of customers treated for campaigns under the proposed targeting policy stated in Eq. 5.2 (*Analytical*). We evaluate the conversion probability estimates provided under the estimation procedures described by the columns *Conversion Model* and *Estimator*. To calculate the expected cost, the analysis includes the model-based approaches discussed above and, for comparison, the expected conversion rate in the population. The conversion rate assumes a constant conversion probability for all customers, which implies a homogeneous treatment cost that is often assumed in studies on cost-sensitive learning. The conversion estimates are combined with two estimation procedures of the treatment effect to calculate the campaign profit. The treatment effect for each customer is either estimated to be the average treatment effect over all customers in the training data, denoted as ATE, or presumed to be estimated perfectly, denoted as *Actual*. The ATE policy makes the simplifying assumption that there exists no heterogeneity in treatment effects and is equivalent to the constant acceptance rate of the treatment assumed in prior studies on cus-

tomer churn (e.g. Lemmens & Gupta, 2017; Verbraken et al., 2012). Beyond the comparison to prior research, the ATE policy provides an estimate of the profit implication of the cost-based targeting alone. Presuming perfect estimation of the ITE is unrealistic in practice, since the true treatment effect is unobservable. As a second comparison, the campaign profit under the actual ITE provides an upper bound on campaign profit that would be achievable by estimation of individual-level cost under optimal performance of the treatment effect model.

Table 5.3 shows that customer-level estimates of the conversion rate provide a more accurate estimate of customer-level costs and translate into higher campaign profit when used in combination with either ATE or CATE estimates. To provide some context for the profit of the models of interest, consider the two simple policies of targeting no or every individual in the population. The Baseline policy, under which no customer is targeted, results in a profit of €46,236. This profit is the result of the natural probability in the customer population to complete a purchase, which we hope to increase with the marketing campaign. Next, consider the average treatment effect of the population and the average conversion rate of the population given treatment. The analytical policy indicates to target all customers given the positive expected average return. The treatment rate of 100% results in a profit of €50,830. The campaign profit defined as the difference between the campaign and no marketing incentive is €4,594.
We now introduce an individual-level targeting policy by estimating the cost of the marketing treatment on the customer level with a response model. Both the two-model and single-model architectures result in a substantial decrease in the fraction of customers treated from 100% to 76%–84%, depending on the estimator. The decrease in treatment ratio is accompanied by an increase in campaign profit between €1,100 and €2,100, again depending on the estimator. This substantial increase of 24-46% in campaign profit compared to universal treatment is the direct result of controlling the expected treatment cost for each customer.
The observed positive impact on profit generalizes to customer-level targeting based on the CATE under treatment effect heterogeneity. A hypothetical targeting policy based on the actual ITE and the average cost results in a campaign profit of €55,493. We again find that campaign profit using customer-level estimates of the treatment cost increase campaign profit by €1,200–€1,900.
Compare now the two-model approach and single model approach with the GBT estimator. The single model GBT results in a campaign profit of €50,830 and €56,696, while the two-model GBT results in a profit of €52,402 and €57,022, for the constant and true treatment estimates respectively. We conclude that the campaign profit under the single-model conversion model is slightly lower than from the campaign profit under the two-model conversion model.

5.5.2 Profit Implications of Causal Hurdle Models

The analysis was so far restricted to the conversion models and the effect of customer-level cost estimation. Considering the probability of each customer to accept the costly marketing incentive directly results in a substantial profit increase. We therefore conclude that the estimate $p(C|X_i, T = 1)$ is sufficiently precise for profitable targeting and that the expected individual cost is practically relevant for customer targeting. The conclusion applies to campaigns consid-

ering heterogeneous treatment effects and population-level estimates of the average treatment effect. In contrast to prior work (e.g. Gubela et al., 2019), our analysis implies that customers with a positive response to treatment can be unprofitable targets due to a high conversion probability after treatment and the associated higher expected treatment cost.

Table 5.4: Quality of model estimates for the conditional average treatment effect

Architecture			Error	
CATE Model	Stages	Estimator	RMSE	TOL
ATE	-	-	2.75	3387.90
Single-Model	Hurdle	GBT	2.37	3384.91
Two-Model	Hurdle	Linear	5.15	3410.78
Two-Model	Hurdle	GBT	1.94	3381.79
Single-Model	One-Stage	GBT	2.77	3387.49
Two-Model	One-Stage	Linear	4.16	3407.13
Two-Model	One-Stage	GBT	1.94	3381.76
DR	One-Stage	Linear	4.11	3406.10
DR	One-Stage	GBT	2.37	3385.45
Actual	-	-	0.00	3374.99

TOL: Transformed Outcome Loss on the observed outcomes
RMSE: Root Mean Squared Error on the simulated treatment effect.

We now consider the estimation of the CATE for the customer-level prediction of marketing effectiveness. We evaluate the quality of the CATE models using statistical indicators and the resulting profit as part of a targeting policy.
Table 5.4 shows the root-mean- squared error (RMSE) of the CATE estimates compared to the simulated treatment effect on profit and the transformed outcome loss (TOL) on the observed outcomes. We include the TOL as a feasible metric when the true treatment effect is not simulated and therefore not known (Athey & Imbens, 2015). To put the results into context, the ATE estimate provides the baseline obtained by a constant estimator, while the actual ITE in the last row provides the lowest obtainable TOL on the data. Kernel density plots showing the distributional fit of the CATE estimates are available in Figure 5.2 in the Appendix.

The linear model is consistently outperformed by the GBT and, on average, ranks below the constant treatment effect estimate. The linear model achieves an RMSE of 5.15, 4.16 and 4.11 within the two-model hurdle and the two-model and doubly-robust one-stage architectures, respectively. The noticeably high RMSE for the two-model hurdle architecture is the result of treatment effect estimates with high absolute value for a small number of observations. The good calibration of the linear model may nevertheless ensure its value within a targeting policy. Under GBT specification, the two-model hurdle architecture compares favorably to the single-model hurdle architecture and models estimating the overall treatment effect directly. The two-model architecture achieves an RMSE of 1.94 for both the hurdle and one-stage model, respectively. The single-model architecture, in comparison, achieves an RMSE of 2.37 and 2.77 for the respective target. The one-stage doubly-robust model with an RMSE of 2.37 performs better than the single-model architecture but worse than the two-model approach.

The results suggest that the single-model approach, which requires the least number of models to be estimated, provides worse estimates of the treatment effect than the two-model or DR models. Analysis of the resulting policy profit will clarify if the gap in estimation precision results in a substantial effect on campaign profit in practice.

Table 5.5: Campaign profit for CATE-based targeting under population average cost estimates

	Architecture				Profit	Fraction Treated
Policy	Stages	CATE Model	Estimator	Conversion Model		
Baseline	-	-	-	-	46,236	0.00
Analytical	-	ATE	-	Conversion Rate	50,830	1.00
Analytical	Hurdle	Single-Model	GBT	Conversion Rate	48,840	0.20
Analytical	Hurdle	Two-Model	Linear	Conversion Rate	54,550	0.66
Analytical	Hurdle	Two-Model	GBT	Conversion Rate	55,590	0.70
Analytical	One-Stage	Single-Model	GBT	Conversion Rate	52,795	0.41
Analytical	One-Stage	Two-Model	Linear	Conversion Rate	54,456	0.66
Analytical	One-Stage	Two-Model	GBT	Conversion Rate	55,146	0.72
Analytical	One-Stage	DR	Linear	Conversion Rate	54,459	0.66
Analytical	One-Stage	DR	GBT	Conversion Rate	54,629	0.83
Analytical	-	Actual	-	Conversion Rate	55,493	0.71

The campaign profit from customer-level targeting provides an interpretable evaluation of the CATE models. Table 5.5 reports the campaign profit resulting from each CATE model in combination with a constant targeting cost derived from the population average conversion probability. When applied within a targeting policy, the conclusions drawn from Table 5.4 are only partially supported.

The linear models are highly profitable when used as part of a targeting policy. With campaign profit of €54,550, €54,456 and €54,459, the linear models are superior to the constant treatment estimate with a profit of €50,830 despite their higher RMSE. The linear specification is, however, dominated by the GBT specification for all architectures except the single-model.

Within the GBT specification, the single-model approach is substantially less profitable than other architectures. The two-model hurdle model, two-model one-stage model and doubly-robust one-stage model show no substantial difference at a campaign profit of €55,590, €55,146 and €54,629, respectively. Campaign profit is substantially worse for the single-model architecture, with a profit of €48,840 for the hurdle model and a profit of €52,795 for the one-stage model. The results confirm that small differences in the precision of the CATE estimates have a practically relevant effect on campaign profit. Despite the hurdle single-model GBT showing a lower RMSE than the ATE baseline in Table 5.4, it underperforms the baseline of uniform treatment by €1,990 when applied for targeting. For all other approaches, we observe a substantial increase in campaign profit under the analytical targeting policy relative to uniform targeting in the range of €3,626–€4,760. With regard to the comparison between the hurdle and one-stage approaches, the results suggest that the two-stage hurdle model results in campaign profit equivalent to that of the one-stage approaches.

5.5.3 Profit Implications of the Proposed Analytical Targeting Policy

The analysis has so far addressed the evaluation of the CATE and conversion estimates separately. We now evaluate the joint impact on campaign profit of the interaction between the proposed treatment and conversion models as part of a targeting policy. Recall that the single- and two-model hurdle models provide an explicit estimate of the conversion probability by design. CATE models that estimate the treatment effect on the profit directly require a separate classification model to predict the conversion rate under treatment.

Table 5.6: Campaign profit for CATE-based targeting under model-based cost estimation

	Architecture				Profit	Fraction Treated
Policy*	Stages	CATE Model	Conversion Model	Estimator		
Baseline	-	-	-	-	46,236	0.00
Analytical	-	ATE	-	Conv. Rate	50,830	1.00
Analytical	Hurdle	Single-Model	-	GBT	54,665	0.53
Analytical	Hurdle	Two-Model	-	Linear	56,172	0.71
Analytical	Hurdle	Two-Model	-	GBT	56,084	0.71
Analytical	One-Stage	Single-Model	Separate	GBT	52,881	0.49
Analytical	One-Stage	Two-Model	Separate	Linear	56,010	0.66
Analytical	One-Stage	Two-Model	Separate	GBT	55,942	0.68
Analytical	One-Stage	DR	Separate	Linear	56,028	0.66
Analytical	One-Stage	DR	Separate	GBT	55,160	0.75
Empirical	Hurdle	Single-Model	-	GBT	53,964	0.78
Empirical	Hurdle	Two-Model	-	Linear	54,940	0.73
Empirical	Hurdle	Two-Model	-	GBT	55,311	0.70
Empirical	One-Stage	Single-Model	-	GBT	54,546	0.69
Empirical	One-Stage	Two-Model	-	Linear	54,269	0.68
Empirical	One-Stage	Two-Model	-	GBT	55,295	0.70
Empirical	One-Stage	DR	-	Linear	54,481	0.67
Empirical	One-Stage	DR	-	GBT	54,791	0.83

Empirical denotes targeting based on the profit-maximizing threshold on the training data.

Table 5.6 reports the campaign profit under the proposed analytical targeting policy and the empirical thresholding policy introduced in Section 5.2. Recall that the analytical policy employs the estimated CATE and conversion probability under treatment to calculate the expected profit from targeting the customer using the decision rule proposed in Eq. 5.2. The empirical policy determines the profit-optimal threshold on the CATE estimates through numeric optimization of the overall campaign profit on the training data. The analytical targeting policy results in a higher campaign profit relative to the baseline for all model architectures and relative to the empirical policy for seven out of eight architectures.

Comparing model architectures, we find that the proposed causal hurdle framework performs at least comparable to the combination of a one-stage treatment effect model with a separate conversion model. All architectures under the analytical and empirical policy increase the campaign profit compared to uniform targeting. Compared to the baseline, the analytical policy increases campaign profit by €2,051–€5,342. The increase in campaign profit by combining estimates

of the CATE and expected cost results in a median additional increase of €1,000 over the treatment-based policy ignoring response-dependent cost reported in Table 5.5. The campaign profit compared to no targeting lies for the hurdle architectures in the range of €6,645–€9,936 and for the one-stage architectures in the range of €7,728–€9,075. Comparing within the two-model architectures, which predict the treatment effect most precisely, we find no substantial difference at a net campaign profit of around €9,750 for the hurdle two-model and the one-stage two-model approach. We also find no substantial difference to the DR approach evaluated as a state-of-the-art competitor one-stage benchmark.
However, the single-model architecture performs substantially worse than the two-model approaches within the one-stage and hurdle architectures. This finding is in line with the lower precision of the treatment effect estimates reported in Table 5.4. We conclude that the proposed two-model hurdle architecture, although not the single-model hurdle architecture, achieves competitive campaign profit to the alternative one-stage, distinct modeling architectures. Despite its disadvantage in estimation, the single-model hurdle architecture improves the effectiveness of the model building process by reducing the number of models that need to be estimated to two compared to the three to five models required by the distinct and two-model hurdle architectures.

Comparing the same model architecture under the analytical and empirical targeting policy, the proposed analytical targeting policy increases campaign profit by an on average €1000, excluding the single-model approach due its weak absolute performance. This result supports the conclusion that the proposed analytical targeting policy increases campaign profit relative to numeric optimization of the decision threshold. Note, however, that the ratio of customer treated by the single-model approaches deviates from the other architectures under the analytical policy, but not under the empirical policy. We interpret these findings as an issue of model calibration for the single-model approach. If either the probability model or the treatment effect model is not well calibrated, expectation calculations will be inaccurate. In this case, empirical thresholding can be an alternative to model recalibration on the level of the policy rather than recalibration of the model estimates. For calibrated models including the GBT when estimated in the two-model architecture, the analytical targeting policy substantially increases policy profit.

We conclude that the proposed analytical decision policy can substantially increase the profitability of targeting models in practice and that the proposed two-model hurdle model architecture is an efficient way to estimate the necessary decision variables in a unified framework.

5.6 Conclusion

We have presented a general analysis of the customer targeting decision problem under different types of variable costs and proposed a causal hurdle framework to estimate the relevant decision variables efficiently. Our results demonstrate that the consideration of treatment cost substantially increases campaign profit when used for customer targeting independent of whether the treatment effect is considered to vary over customers.

While customer targeting based on expected profit has been used to optimize campaigns, previous analytical frameworks do not include marketing incentives that are conditioned on a profitable customer response, e.g. a retention offer or voucher. We identify these common marketing incentives as a type of stochastic variable cost. Our formal analysis of the targeting decision problem under customer response-dependent costs shows that estimating the expected cost requires an estimate of the customer response conditional on treatment. A central result to the customer targeting literature is that profit-optimal targeting often requires modeling the effect of the marketing treatment and the net customer response under treatment.

In order to estimate the treatment effect and response efficiently, we propose a framework for joint estimation. Our causal hurdle model combines a hurdle model for customer choice with methods for causal inference. The proposed approach is feasible with the single-model and two-model approaches for the estimation of conditional treatment effects. We find that the causal hurdle model under the two-model specification achieves competitive campaign profit on a coupon targeting campaign in an e-commerce setting, while streamlining model building.

With the increasing relevance of digital marketing and the associated increase in marketing incentives with low targeting-dependent and high response-dependent variable costs, our results are highly relevant for practitioners. We further expect the development of efficient approaches for the estimation of flexible hurdle models and the application of our decision analysis to other applications with stochastic costs as fruitful areas for future research.

Bibliography

Ascarza, E. (2018). Retention futility: Targeting high risk customers might be ineffective. *Journal of Marketing Research*, *55*(1). https://doi.org/10.1509/jmr.16.0163

Ascarza, E., Ebbes, P., Netzer, O., & Danielson, M. (2017). Beyond the target customer: Social effects of customer relationship management campaigns. *Journal of Marketing Research*, *54*(3), 347–363. https://doi.org/10.1509/jmr.15.0442

Ascarza, E., & Hardie, B. G. S. (2013). A joint model of usage and churn in contractual settings. *Marketing Science*, *32*(4), 570–590. https://doi.org/10.1287/mksc.2013.0786

Athey, S., & Imbens, G. W. (2015). *Machine Learning Methods for Estimating Heterogeneous Causal Effects*.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178.

Athey, S., & Wager, S. (2017). Efficient Policy Learning. *arXiv preprint*, arXiv:1702.02896.

Backiel, A., Baesens, B., & Claeskens, G. (2016). Predicting time-to-churn of prepaid mobile telephone customers using social network analysis. *Journal of the Operational Research Society*, *67*(9). https://doi.org/10.1057/jors.2016.8

Baumann, A., Haupt, J., Gebert, F., & Lessmann, S. (2019). The price of privacy: An evaluation of the economic value of collecting clickstream data. *Business & Information Systems Engineering*, *61*(4), 413–431. https://doi.org/10.1007/s12599-018-0528-2

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge, Cambridge University Press.

Devriendt, F., Berrevoets, J., & Verbeke, W. (2019). Why you should stop predicting customer churn and start using uplift models. *Information Sciences*, *In Press*. https://doi.org/10.1016/j.ins.2019.12.075

Devriendt, F., Moldovan, D., & Verbeke, W. (2018). A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. *Big Data*, *6*(1), 13–41. https://doi.org/10.1089/big.2017.0104

Diemert, E., Betlei, A., Renaudin, C., & Amini, M.-R. (2018). A Large Scale Benchmark for Uplift Modeling, In *Proceedings of the AdKDD and TargetAd Workshop, KDD*, London, United Kingdom, ACM.

Donkers, B., Paap, R., Jonker, J.-J., & Franses, P. H. (2006). Deriving target selection rules from endogenously selected samples. *Journal of Applied Econometrics*, *21*(5), 549–562. https://doi.org/10.1002/jae.858

Gubela, R. M., Bequé, A., Gebert, F., & Lessmann, S. (2019). Conversion uplift in e-commerce: A systematic benchmark of modeling strategies. *International Journal of Information Technology & Decision Making*, *18*(3), 747–791.

Gubela, R. M., Lessmann, S., & Jaroszewicz, S. (2020). Response transformation and profit decomposition for revenue uplift modeling. *European Journal of Operational Research*, *283*(2), 647–661. https://doi.org/10.1016/j.ejor.2019.11.030

Hansotia, B., & Rukstales, B. (2002). Incremental value modeling. *Journal of Interactive Marketing*, *16*(3), 35–46. https://doi.org/10.1002/dir.10035

Hill, J. L. (2011). Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240. https://doi.org/10.1198/jcgs.2010.08162

Hitsch, G. J., & Misra, S. (2018). Heterogeneous Treatment Effects and Optimal Targeting Policy Evaluation. *SSRN*.

Hruschka, H. (2010). Considering endogeneity for optimal catalog allocation in direct marketing. *European Journal of Operational Research*, *206*(1), 239–247.

Knaus, M. C., Lechner, M., & Strittmatter, A. (2019). Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence. *IZA Discussion Paper*, *12039*.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165.

Lemmens, A., & Gupta, S. (2017). Managing Churn to Maximize Profits. *Social Science Research Network*, 2964906. https://doi.org/10.2139/ssrn.2964906

Lessmann, S., Haupt, J., Coussement, K., & De Bock, K. W. (2019). Targeting customers for profit: An ensemble learning framework to support marketing decision making. *Information Sciences*, *In Press*. https://doi.org/10.1016/j.ins.2019.05.027

Ling, C. X., & Li, C. (1998). Data Mining for Direct Marketing: Problems and Solutions, In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, Menlo Park, US, AAAI Press.

Mitrović, S., Baesens, B., Lemahieu, W., & De Weerdt, J. (2018). On the operational efficiency of different feature types for telco churn prediction. *European Journal of Operational Research*, *267*(3), 1141–1155. https://doi.org/10.1016/j.ejor.2017.12.015

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research*, *43*(2), 204–211. https://doi.org/10.1509/jmkr.43.2.204

Nie, X., & Wager, S. (2017). Quasi-Oracle Estimation of Heterogeneous Treatment Effects. *arXiv e-prints*, 1712.04912.

Olson, D. L., & Chae, B. (2012). Direct marketing decision support through predictive customer response modeling. *Decision Support Systems*, *54*(1), 443–451. https://doi.org/10.1016/j.dss.2012.06.005

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., & Tibshirani, R. (2018). Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine*, *37*(11), 1767–1787. https://doi.org/10.1002/sim.7623

Radcliffe, N. J. (2007). Using control groups to target on predicted lift: Building and assessing uplift models. *Direct Marketing Analytics Journal*, 14–21.

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, *90*(429), 122–129. https://doi.org/10.1080/01621459.1995.10476494

Rzepakowski, P., & Jaroszewicz, S. (2012). Uplift modeling in direct marketing. *Journal of Telecommunications and Information Technology*, *2*, 43–50.

Sahni, N. S., Zou, D., & Chintagunta, P. K. (2016). Do targeted discount offers serve as advertising? Evidence from 70 field experiments. *Management Science*, *63*(8), 2688–2705. https://doi.org/10.1287/mnsc.2016.2450

Shi, C., Blei, D. M., & Veitch, V. (2019). Adapting Neural Networks for the Estimation of Treatment Effects. *arXiv preprint*arxiv 1906.02120, arXiv:1906.02120.

Statista. (2017). *Advertising Spending in the Catalog, Mail-order Houses Industry in the United States* (tech. rep.).

Statista. (2019). *eCommerce* (tech. rep.).

Van Diepen, M., Donkers, B., & Franses, P. H. (2009). Dynamic and competitive effects of direct mailings: A charitable giving application. *Journal of Marketing Research*, *46*(1), 120–133. https://doi.org/10.1509/jmkr.46.1.120

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, *218*(1), 211–229. https://doi.org/10.1016/j.ejor.2011.09.031

Verbraken, T., Verbeke, W., & Baesens, B. (2012). A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE Transactions on Knowledge and Data Engineering*, *25*(5), 961–973. https://doi.org/10.1109/TKDE.2012.50

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N. H., & Gallego, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from

health care databases. *Statistics in Medicine*, *37*, 3309–3324. https://doi.org/10.1002/sim.7820

Zhang, X. A., Kumar, V., & Cosguner, K. (2017). Dynamically managing a profitable email marketing program. *Journal of Marketing Research*. https://doi.org/10.1509/jmr.16.0210

5.A Relation to Previous Formulations of Churn Campaign Profit

A popular definition of the profit of a customer retention campaign (e.g. Devriendt et al., 2019; Lemmens & Gupta, 2017; Verbeke et al., 2012) is given by Neslin et al. (2006):

$$\Pi = N\alpha\left[\beta\gamma(V-\delta-c)+\beta(1-\gamma)(-c)+(1-\beta)(-\delta-c)\right]-A$$

with
N: Number of customers
α: Ratio of customers targeted
V: The value of the customer to the company, CLV in their original notation
β: Fraction of (targeted) customers who would churn
γ: Fraction of (targeted) customers who decide to remain when receiving the marketing incentive
δ: The cost of the marketing incentive if it is accepted
c: The cost of contacting the customer with the marketing incentive
A: The fixed cost of running the retention campaign

The number of customers targeted by the campaign and the fixed costs are relevant to calculate the overall campaign profit, but do not affect the targeting decision for a single customer. The profit estimate relevant for customer targeting is thus the part in square brackets:

$$\pi_i = \beta_i\gamma_i(V-\delta)+\beta_i(1-\gamma_i)(-c)+(1-\beta_i)(-\delta-c)$$

We will show that this expression is equivalent to the proposed decision policy (Eq. 5.2) under restrictive assumptions. Using the additive property of the probabilities β_i and $(1-\beta_i)$ and γ_i and $(1-\gamma_i)$, we can summarize the terms:

$$\begin{aligned}\pi_i &= \beta_i\gamma_i(V)+\beta_i\gamma_i(-\delta)+(1-\beta_i)(-\delta)+\beta_i(-c)+(1-\beta_i)(-c)\\ &= \beta_i\gamma_i V+\beta_i\gamma_i(-\delta)+(1-\beta_i)-\delta-c\\ &= \beta_i\gamma_i V-\delta(\beta_i\gamma_i+1-\beta_i)-c\\ &= \beta_i\gamma_i V-\delta(1-\beta_i(1-\gamma_i))-c\end{aligned}$$

We will target a customer if the profit is positive, i.e.

$$\beta_i\gamma_i V-(1-\beta_i(1-\gamma_i))\delta-c>0 \tag{5.12}$$

In Eq. 5.2, we propose the decision rule

$$p_i(1)(V(1)-\delta)-c>p_i(0)\cdot V(0)$$

Assuming that the value of the customer is not influenced by the marketing incentive $V(1)=$

$V(0) = V$ allows us the rearrange the inequality to

$$(p_i(1) - p_i(0))V - p_i(1)\delta - c > 0 \tag{5.13}$$

Eq. 5.12 and Eq. 5.13 are equivalent if the following equalities hold:

$$\begin{aligned} p_i(1) &= (1 - \beta_i(1 - \gamma_i)) \\ p_i(1) - p_i(0) &= \beta_i\gamma_i \end{aligned}$$

In words, we require $p(1)$ to be the complement to the probability for a customer to plan to churn and churn even when offered the treatment. The complimentary event is for a customer not to plan to churn or to plan to churn but remain after treatment; or simply, the probability of the customer to stay when given treatment.

We further require $p(1) - p(0)$ to be the probability of a customer to plan to churn and to not churn when offered the treatment. As $\beta_i \cdot \gamma_i \in [0; 1]$, this equality holds under the assumption that the treatment effect is strictly positive, i.e. $p(1) - p(0) \in [0; 1]$. However, we know that the treatment effect on the response probability, $p(1) - p(0)$, is in principle bounded in $[-1, 1]$ and that negative effects are a critical issue in churn campaigns in practice (Ascarza, 2018). Under the previous campaign profit formulation, we see that $\beta_i\gamma_i = 0$ if either β_i or γ_i or both are zero. In words, the campaign has no effect if no customers consider to churn or no customers accept the marketing incentive when offered. This conflicts with the observation that when no customers plan to churn, the campaign may have a net negative effect by priming inattentive customers to churn. Specifically, the shortcoming of the customer profit proposed by Neslin et al. (2006) is that it implicitly assumes a positive treatment effect by restricting the action space of the customer to $\gamma \in \{\text{Accept treatment}, \text{Disregard treatment}\}$.

We conclude that the proposed decision framework is a generalization of Neslin et al. (2006)'s campaign profit function to cases where a customer may react adversely to the treatment. As an alternative formulation to calculate the overall churn campaign profit, we propose for the general case:

$$\Pi = \sum_{i \in N} \{T_i \left[(p_i(1) - p_i(0))V_i - p_i(1)\delta - c\right]\} - A$$

In cases with no or little variation in customer sensitivity to the marketing treatment and a constant customer lifetime value, the churn campaign profit can be simplified to:

$$\Pi = N\alpha \left[\hat{\tau}_{ATE}V - p(1)\delta - c\right] - A$$

5.B Additional Evaluation Results

Table 5.7 shows the quality of predictions for the conversion probability conditional on treatment. Recall that the single-model hurdle model includes the treatment indicator as a covariate into the model. The two-model hurdle model estimates four separate models, one of which predicts the conversion probability within the treatment group. Note that the default approach, which separates treatment effect estimation and conversion prediction, also requires the estimation of an identical conversion model. This is the redundancy that the proposed causal hurdle framework avoids. We find no substantial difference in the area-under-the-ROC-curve (ROC-AUC) or the Brier score, which indicates model calibration.

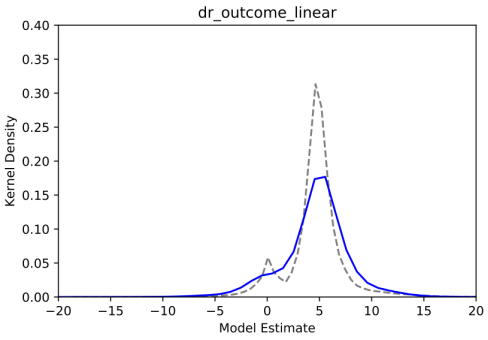
Table 5.7: Quality of model estimates for the prediction of conversion under treatment

Architecture				
Stages	Specification	Estimator	ROC-AUC	Brier Score
Hurdle/Distinct	Two-Model	Linear	0.636	0.103
Hurdle/Distinct	Two-Model	GBT	0.640	0.102
Hurdle Model	Single-Model	GBT	0.636	0.102

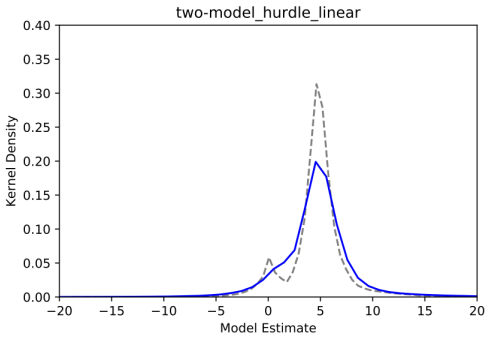
Figure 5.2 depicts the kernel density plot for the treatment estimation approaches and the GBT specification. We combine the out-of-sample estimates for each iteration of the cross-validation procedure to obtain out-of-sample estimates for the full dataset. The dotted line shows the kernel density of the actual ITE.

We observe that no approach fully captures the minor mode of the distribution to the left. The hurdle single-model GBT approach in addition shows a slight shift from the major mode of the distribution that relates to the worse precision reported in Table 5.4.

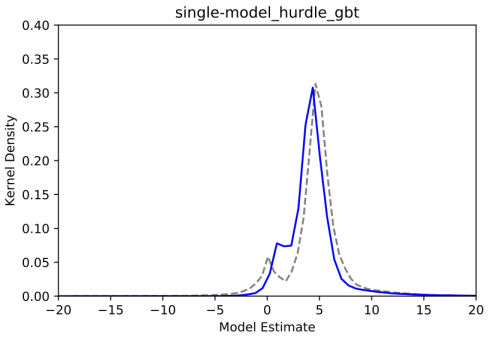
The support of the linear model specifications extends beyond the actual range of the simulated treatment effects and beyond the range shown in the figure. For a small set of observations, we observe predicted treatment effects beyond the range [-100;100] that explain the high statistical error reported in 5.4. For the remaining observations, we observe a reasonable fit to the actual treatment effect distribution. The general fit explains the profitability of the linear specification for the targeting policy as observations with weak support, for which linear extrapolation fails, by definition make up only a minority of cases in the data.



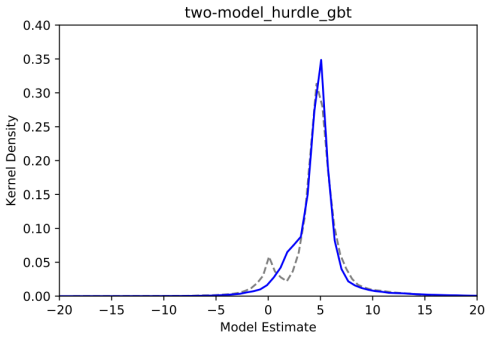
(a) One-Stage DR Linear



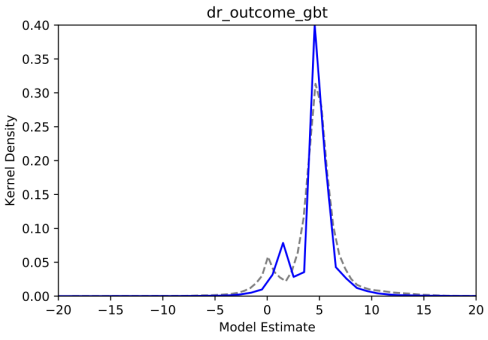
(b) Hurdle Two-Model Linear



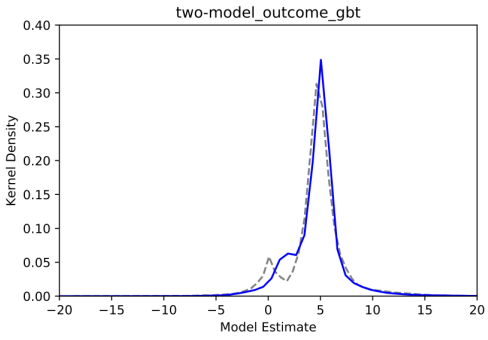
(c) Hurdle Single-Model GBT



(d) Hurdle Two-Model GBT



(e) One-Stage Doubly-Robust GBT



(f) One-Stage Two-Model GBT

Figure 5.2: Kernel density plot of the CATE on the outcome as estimated by the hurdle (top rows) and one-stage models (bottom). The dotted line shows the actual individual treatment effect.

Chapter 6

Supervised Randomization in Controlled Experiments

Publication
Haupt, J., Jacob, D., Gubela, R., & Lessmann, S. (2019). Affordable Uplift: Supervised Randomization in Controlled Experiments. Proceedings of the 40th International Conference on Information Systems (ICIS).

Abstract
Customer scoring models are the core of scalable direct marketing. Uplift models provide an estimate of the incremental benefit from a treatment that is used for operational decision-making. Training and monitoring of uplift models require experimental data. However, the collection of data under randomized treatment assignment is costly, since random targeting deviates from an established targeting policy. To increase the cost-efficiency of experimentation and facilitate frequent data collection and model training, we introduce *supervised randomization*. It is a novel approach that integrates existing scoring models into randomized trials to target relevant customers, while ensuring consistent estimates of treatment effects through correction for active sample selection. An empirical Monte Carlo study shows that data collection under supervised randomization is cost-efficient, while downstream uplift models perform competitively.

6.1 Introduction

Direct marketing plays a key role in consumer markets. The continuous growth of e-commerce, accounting for 1.8 trillion Euros globally in 2019 (Statista, 2019), is accompanied by simultaneous growth in online and email advertising. Spending on traditional print advertising like catalog marketing has shown a similar growth (Statista, 2017). At the core of scalable direct marketing, campaign analysts employ models to predict future customer behavior and target responsive clients (Olson et al., 2012).

For example, a decision tree could be trained to predict the probability for a customer to purchase in the next week based on known characteristics. The expected behavior of the customer could then be used to inform operational decision-making in that customers with a probability below average are targeted with an incentive. However, the predictive model is agnostic to the marketing policy, the overall effectiveness of the marketing action and the effect of the marketing action on individual customers. Outcome models provide an estimate of

customer behavior, but fail to provide an estimate of the potential change in customer behavior, which is the goal of marketing intervention.

A growing research stream advocates that the decision which customers to target should be addressed directly through causal inference in the form of uplift models (Devriendt et al., 2018; Gubela et al., 2017). Instead of predicting customer behavior, uplift models estimate the causal effect of a marketing action on an individual customer given their characteristics. In the above example, an uplift tree could be trained to predict the increase in probability for a customer to conduct a purchase in the next week if a catalog was sent. Uplift models thus provide an estimate of the incremental benefit from the marketing treatment, which can explicitly be used as a direct criterion for operational decisions by comparing it to the incremental cost. Conceptually, uplift models align with the actual decision problem of choosing the action with the highest incremental gain for each customer.

Uplift models are trained on experimental data and estimate the treatment effect by comparing the observed behavior of a group of individuals who have received the treatment, the treatment group, and a distinct group of individuals who have not received the treatment, the control group. Similarly, experimental data is required to evaluate the performance of uplift models (Radcliffe, 2007). In contrast, non-causal models of customer behavior are trained and evaluated on customers of which all or none have received the treatment. Collecting experimental data in randomized experiments is well established in practice in the form of A/B tests. Although used to evaluate the gross benefit of campaigns, A/B tests are not commonly used for uplift modeling to estimate individualized treatment effects (Ascarza, 2018).

During experiments, random assignment of individuals to either the treatment or control group is crucial to train unbiased uplift models. However, data collection through randomized experiments is costly, since random targeting withholds marketing spending on some customers that would be targeted under the established targeting policy and applies spending on customers that would otherwise not be targeted. The deployment of uplift models exacerbates data collection costs since decision support systems typically require continuous or frequent evaluation and occasional retraining on recent observations, which in turn require fresh experimental data.

We propose a novel approach for the collection of experimental data for uplift modeling based on the combination of cost-optimized randomization at the time of data collection and selection bias correction during model building, which we refer to as *supervised randomization*. In a nutshell, supervised randomization introduces a stochastic component to the existing targeting model and extends the standard experimental design with full randomization by considering customers that are rejected by the targeting policy.
Our contribution is two-fold. First, we show that supervised randomization can be used to integrate existing scoring models into randomized trials. The integration of existing scoring models into group assignment increases the cost-efficiency of experimentation and facilitates continuous data collection during regular business operation. Continuous data collection is critical for non-disruptive experimentation, monitoring the performance of uplift models under deployment and recurring model training. Facilitating model training and monitoring has the

additional benefit to improve the acceptance of causal models by management and stakeholders. Second, we introduce inverse probability weighting and doubly robust estimation as methods to control for biased treatment assignment to the uplift literature. Uplift models have so far relied on the assumption of data collected under full or imbalanced randomization in randomized controlled trials. We show that recent advances in the econometrics literature extend the applicability of uplift models to cases with non-standard treatment assignment. The bias-corrected uplift models are shown to perform competitively on simulated data.

6.2 Background

Consider a marketing action applied to an individual user i as a treatment intended to change an observed outcome Y_i. Let $D_i \in 0, 1$ be an indicator if the individual has been treated and denote the outcome with and without treatment as Y_i^1 and Y_i^0, respectively. Then the individual treatment effect is the incremental gain caused by a marketing action $Y_i^1 - Y_i^0$. Because a customer either does or does not receive the marketing action, the actual treatment effect is not observable. We can, however, estimate the treatment effect on the population or on the individual level. We denote the average campaign uplift as average treatment effect (ATE) and the customer-level uplift $\tau = E[Y_i^1 - Y_i^0 | x = X_i]$ as individualized treatment effect (ITE) (Knaus et al., 2019; Powers et al., 2018), sometimes alternatively denoted conditional average treatment effect (CATE) in the econometrics literature. Furthermore, we refer to a model used to estimate the outcome Y_i as outcome model and a model used to estimate the treatment effect τ as a causal model, as a more general alternative to the term uplift model. The operational decision problem posed in uplift modeling is to decide if an individual customer should receive the marketing treatment. The decision is automated through the targeting policy, a mapping from the estimated ITE to the binary decision of whether to treat the individual.

Three assumptions are needed for causal inference following the potential outcome theorem (Rosenbaum & Rubin, 1983). First, the *Stable Unit Treatment Value Assumption (SUTVA)* guarantees that the potential outcome of a customer is unaffected by changes in the treatment assignment of other customers. This assumption may be violated when treatment effects propagate through the social network of customers (Ascarza et al., 2017). In settings of low value or low involvement products, research on treatment effects in marketing typically assumes that no interaction takes place (Hitsch & Misra, 2018) or that the indirect effect of treatment on other customers is at least substantially smaller than the direct effect of the treatment (Imbens & Wooldridge, 2009).

The second assumption is *conditional unconfoundedness*, i.e. the independence between the potential outcomes and the treatment assignment given the observed covariates (X) (Rosenbaum & Rubin, 1983).

$$\left(Y_i^1, Y_i^0\right) \perp\!\!\!\perp D_i | X_i = x. \tag{6.1}$$

The third assumption, called *overlap*, guarantees that for all $x \in supp(X_i)$ the probability to

receive treatment $e(x) = \mathrm{P}(D = 1|X_i = x)$ is bounded away from 0 and 1:

$$0 < e(x) < 1. \tag{6.2}$$

When the treatment assignment process is under the control of the experiments as in the customer targeting setting, conditional independence and overlap can be ensured by design through fully randomizing treatment assignment with treatment probability $e(x) = e \in (0; 1)$. Randomized experiments assign individuals at random to one of at least two conditions, where each condition entails a specific treatment. In controlled experiments, one condition is the control condition in which individuals receive no treatment. In combination, randomized controlled trials (RCT) are the gold standard of data collection for causal inference. We refer to uniform treatment assignment as *full randomization*. Supervised randomization provides a framework that preserves the advantage of the randomized experimental design but allows some control over the probability of treatment assignment on the individual level.

6.3 Literature Review

The unbiased training of causal models and targeting policies requires data that fulfills the assumptions of the potential outcome theorem. In addition, the unbiased evaluation of causal models and policies also requires experimental data and metrics developed for counterfactual prediction (Hitsch & Misra, 2018; Radcliffe, 2007). Violation of the unconfoundedness (Eq. 6.1) and overlap assumptions (Eq. 6.2) in observational studies cannot be substituted by collecting more data in the form of more covariates or more observations (Gordon et al., 2019). Randomized experiments are thus considered a prerequisite to uplift modeling. However, the design and costs of RCT are often not discussed in the literature. We aim to fill this gap by proposing a more efficient design for randomized experiments. We first summarize recent developments in causal machine learning, related research on efficient experimental design and methods to correct for treatment assignment in observational studies.

Causal machine learning methods can be divided into direct and indirect approaches. Direct estimation algorithms construct a feasible loss to estimate a model for the ITE. Indirect approaches model the expected customer response conditional on the treatment group and estimate the ITE as the difference between expected responses. This study employs a robust, indirect two-model logistic regression and a state-of-the-art, direct causal forest for the empirical comparison and provides a discussion of these models below. For an in-depth discussion and benchmark of recent methods for ITE estimation see (Knaus et al., 2019; Künzel et al., 2019; Powers et al., 2018).

Indirect approaches estimate the treatment effect via estimating the response with and without treatment using common statistical learners. The two-model approach (Radcliffe, 2007), or k-model approach in settings with more than one treatment, estimates a separate model for the outcome in the treatment group and control group data and estimates the ITE as the difference between the predicted outcomes. The two-model approach is flexible with regard to the underlying outcome models. K-nearest neighbors learners (Gubela et al., 2019) and deep

neural networks (Farrell et al., 2018) have demonstrated promising model performance in the two-model framework. While recent research advocates discretizing the outcome variable to use classification models in continuous settings (Gubela et al., 2017), the two-model approach extends naturally to both categorical and continuous outcomes. This facilitates the use of classification and regression models to forecast, for example, purchase completion or customer spending, respectively.

A number of well-known machine learning algorithms have been extended to estimate the ITE directly without the need to model the customer response (Lo, 2002; Zaniewicz & Jaroszewicz, 2013). Note that the average treatment effect within a subgroup provides a useful estimate of the treatment effect for individuals within that subgroup. Hence, algorithms that split the data into groups to calculate estimates on the subset are inherently applicable to causal modeling and modifications of the k-nearest neighbor estimator (Hitsch & Misra, 2018) and tree-based models (Athey & Imbens, 2016; Rzepakowski & Jaroszewicz, 2012) have been applied to estimate individualized treatment effects. Causal tree models modify the Classification and Regression Tree by a splitting criterion maximizing the expected variance in treatment effects between leaves (Athey & Imbens, 2016). Within each terminal node conditional on the covariate splitting, the conditional average treatment effect can be estimated and provides an ITE for the observations falling into that node.
Causal trees can be combined into ensembles through bagging or boosting. (Powers et al., 2018) propose a gradient-boosted ensemble of causal trees and an algorithm using multivariate adaptive regression splines. Causal forests are similarly flexible models and have been shown to be consistent and asymptotically normal for a fixed covariate space (Athey et al., 2019).

Both direct and indirect approaches to ITE estimation share the need for experimental data. The collection of experimental data has not been explicitly explored in the uplift literature. However, concerns over the organizational difficulty and the opportunity cost of running randomized controlled trials have led to research on the optimal use of available data and efficient experimental design in related fields.
A popular strategy for the evaluation of multiple targeting policies is to avoid experimentation for each candidate policy and instead to estimate each policy's performance using one existing, fully randomized experiment. The cost-efficient evaluation is possible through extrapolation from observations where the policy recommendation matches the observed random treatment assignment, weighted to match the actual population (Athey & Wager, 2017; Hitsch & Misra, 2018; Swaminathan & Joachims, 2015). This evaluation strategy requires existing experimental data, while our goal is to decrease the cost of collecting experimental data through efficient randomization. Approaches to efficient evaluation and efficient randomization are therefore complementary.

The experimental design of previous studies indicates awareness of data collection costs. Table 6.1 shows the marketing goal, data accessibility, number of observations and the imbalance between treatment and control group sizes of experimental campaigns in customer targeting applications. The large number of observations in recent studies is unsurprising since common

technologies in e-commerce settings (e.g., web cookies) facilitate the collection of large data volumes of customer interactions in online shops Diemert et al. (2018). However, large-scale experimentation imposes substantial costs by randomly withholding profitable treatment for a sizeable control group. We reason that large experiment sizes indicate that companies perceive potential gains from causal modeling and are willing to collect data on a sufficient scale. Since the costs of experimentation are a result of the randomization of treatment assignment, we propose that supervised randomization can lead to cost reductions that are economically relevant in practice given the scale of experimentation. The savings potential increases with the targeting cost and will thus be most effective for catalog or telephone marketing, where resource-intensive treatments drive cost, and in customer churn management, where targeting customers may increase awareness of contract expiration and induce churn in otherwise passive customers.

We further observe that 10 of 11 datasets show a substantial difference in size between the treatment and control group, which we denote *imbalanced full randomization*. The imbalance implies that companies assign customers to the treatment group with probabilities 2–17 times higher than assignment to the control group. The observed imbalanced experimental design is more efficient than equal assignment to treatment and control group when the marketing action is expected to be profitable on average and treatment is the dominant targeting strategy. Companies are thus conducting active cost management of random experiments based on an assessment of overall treatment effectiveness. Our approach follows the same motivation, but extends cost management to the individual level based on an assessment of the individual treatment effectiveness.

Table 6.1: Randomized treatment data in marketing

Application	Source	Obs. (in 1000)	T:R Ratio*
Direct mail in office supplies	Kane et al. (2014)	460	17:1
Mail promotion	B. J. Hansotia and Rukstales (2002)	550	10:1
Cross-selling mail in insurance	Guelman (2014)	34,370	9:1
MSN subscription	Chickering and Heckerman (2000)	110	9:1
Criteo online advertising campaign	Diemert et al. (2018)	29,106	7:1
Direct mail in financial services	Kane et al. (2014)	1,144	5:1
Simulation study	Lo (2002)	100,000	4:1
Customer retention mail in insurance	Guelman et al. (2015)	12	2:1
Catalog marketing	Hitsch and Misra (2018)	441	2:1
E-mail promotion in merchandising	Hillstrom (2008)	64	2:1
E-mail promotion in holiday marketing	B. Hansotia and Rukstales (2002)	282	1:1

*Treatment:Control Ratio

The design of randomization on the individual level is more thoroughly discussed in the medical literature (Schulz & Grimes, 2002). On the one hand, administering a pharmaceutical to a random patient can induce severe health issues, so randomized trials pose a risk for patient health. On the other hand, new treatment may prove to be a substantial improvement over comparative options, so that withholding treatment can be seen as suboptimal care. The latter concern for optimal treatment of patients has motivated research on adaptive randomization procedures,

where patients are more likely to be assigned to treatment for which positive outcomes have been previously observed over the course of the study (Lachin et al., 1988; Rosenberger & Lachin, 1993). Response-adaptive randomization in medical trials differs from our approach in that we use a scoring model to adjust treatment probability conditional on customer characteristics rather than observing treatment outcomes during the trial.

The trade-off between collecting more data to improve the scoring model and applying an existing treatment policy deterministically corresponds to the exploration-exploitation problem in reinforcement learning and multi-armed bandit approaches. Supervised randomization is related to the ε-greedy algorithm (Schwartz et al., 2017), extended by heterogeneous exploration probabilities $\varepsilon_i = 1 - e(x)$. In comparison to upper confidence bound sampling or Thompson sampling (Schwartz et al., 2017) which favor exploration of uncertain predictions, supervised randomization favors exploration close to the decision boundary of the policy and facilitates straightforward logging of the true treatment probability.
This study considers supervised randomization for continuous evaluation and periodical updating of treatment effect models. We do not adapt the scoring model and conditional treatment probabilities during the duration of the experiment as opposed to online learning of the treatment effect model under reinforcement learning. We leave a more in-depth comparison for future research.

Supervised randomization introduces dependency between the covariates and treatment assignment in the data as a side effect of adjusting treatment probabilities on the individual level. This violates the conditional independence assumption and without correction would lead to biased treatment estimates known as selection bias. An intuitive interpretation is that selection bias is due to the covariates being non-identically distributed between the treatment and control group because group assignment is itself based on the observed covariates. Statistical analysis of the average or individualized treatment effect on data that violates the unconfoundedness assumption thus requires correction for the effect of the covariates on the individual probability to receive treatment. To the best of our knowledge, methods to systematically correct for selection bias have not yet been studied in the uplift community. Instead, research on uplift modeling assumes the feasibility of randomized control trials where there is no such selection bias by design, but ignores the associated costs of data collection in practice.
However, selection bias corrections are well-understood and common for research using observational data, where random treatment assignment is not ethical or feasible. The most common technique for selection bias correction are the inverse probability weighting estimator (Horvitz & Thompson, 1952) (IPW) and its extension to the doubly robust (DR) estimator (Robins et al., 1994). Selection bias correction has recently been integrated into popular ITE estimators, which are applied in the observational studies prevalent in economics research (Athey et al., 2019; Künzel et al., 2019).

Within the field of information systems, the IPW correction is used in observational studies (Marco Caliendo et al., 2012) and research on recommender systems and reinforcement learning. In recommendation settings, explicit and some forms of implicit feedback can be understood as

the outcome of a non-randomized experiment, where users evaluate a subset of items that they select based on their preferences. The customer feedback can be corrected by weighting the feedback by the probability of a customer to interact with an item before rating it (Schnabel et al., 2016).
Reinforcement learning research makes regular use of existing log data, which is cost-efficient to collect and available at scale, but not fully randomized. Under some conditions, unbiased training or evaluation of a learning algorithm is possible using IPW to correct for the treatment policy at the time of data collection (Swaminathan & Joachims, 2015). Interestingly, if treatment under the existing policy is stochastic with a known probability and the treatment probability is logged, the resulting process can be seen as an online version of the supervised randomization process.

For observational data, the treatment probability used to correct for the selection bias is unknown and IPW thus follows a two-step process. In the first step, the treatment probabilities used for IPW correction are estimated from the observed data. In the second step, the treatment probability estimates are used to correct subsequent estimates of treatment effects. The approach proposed here is substantially different from the common applications of IPW in the first step. Under supervised randomization, the true treatment probability for each observation is actively controlled and thus known, eliminating the need for estimation and the potential for estimation error through unobserved variable bias or model misspecification.

6.4 Efficiently Randomized Experimental Design

Within this study, we take a holistic view of the causal modeling process emphasizing the interaction between data collection and model building. Data collection through a randomized controlled trial is a necessary part of the causal modeling process. To collect RCT data, the established targeting policy is temporarily replaced by randomized treatment assignment. However, the replacement of an efficient targeting strategy by a random assignment has negative side effects in practice, even when restricted to a subset of customers. First, randomized treatment assignment carries opportunity costs resulting from targeting the wrong customers. Compared to an existing effective targeting strategy, profitable customers are less likely to be targeted while less profitable customers are more likely to be targeted. Second, customers may misconceive data collection periods as a decrease in service or advertising quality. Since customers are not informed about the temporary replacement of the targeting model, they will attribute the random treatment assignment to the targeting efforts of the company.

Instead of replacing the established targeting policy with randomized treatment allocation, we propose to introduce a stochastic component to the existing targeting model as *supervised randomization*. Under supervised randomization, treatment assignment is largely driven by the effective targeting model but sufficiently randomized to allow the estimation and evaluation of causal models. Embedding the existing targeting model into the experimental design has three merits.
First, supervised randomization increases the return on treatment during experimentation when

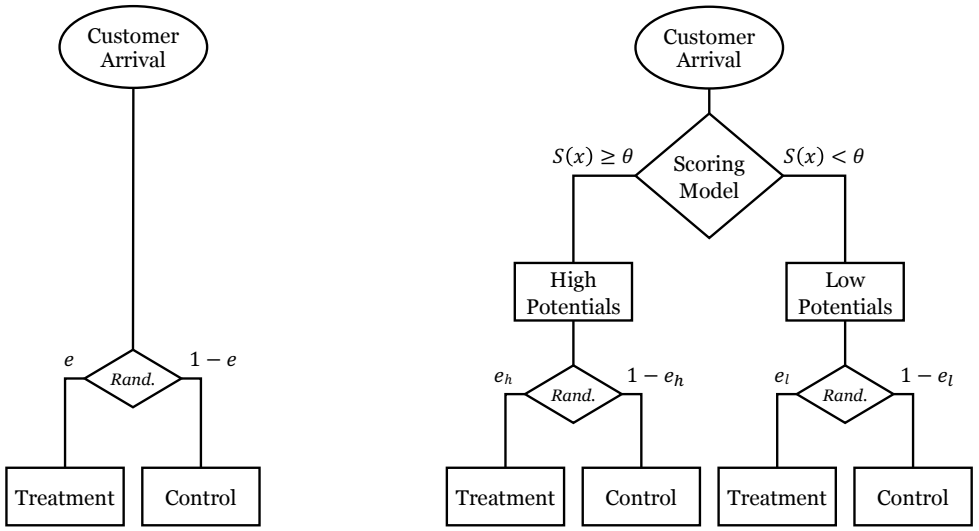
compared to full randomization. Supervised randomization allows us to actively decrease the cost of running randomization experiments by treating profitable customers identified by the targeting model with a higher probability than customers identified as less profitable by the scoring model.
Second, supervised randomization with conservative propensity mapping could facilitate continuous experimentation. Since both training and evaluation of causal models require experimental data, regular repetition of experimental data collection is necessary when causal models are deployed. Continuous experimentation could reduce variability in service from the perspective of the customer and streamline data collection for the company by avoiding interference with operation to run experiments and reducing the need to justify and approve the need of data collection at intervals. These goals are shared with reinforcement or bandit learning with the important difference that our approach facilitates model estimation through standard machine learning or uplift methodology.

6.4.1 Supervised Randomization

Supervised randomization introduces heterogeneity in treatment probabilities into a randomized controlled experiment. We discuss the proposed approach as an extension to A/B testing through the introduction of a targeting model in several steps. We describe the process for one treatment and one control group in the online context where customers arrive in sequence, but note that the same process extends to more than one treatment group and other static settings. A/B testing for treatment evaluation is an instance of randomized controlled experiments with a single treatment. Each arriving customer is randomly assigned to the treatment or control group. The probability to receive treatment is identical for all customers $e(x) = e$ with the probability for assignment to the control group $1 - e$. The probability of treatment assignment can be equal $e = 1 - e = 0.5$ or imbalanced towards the preferred strategy for $e \in (0; 1)$. As discussed in the literature review, imbalanced probabilities are used to control the costs of the experiment in practice. For the case of multiple treatments, a different probability can be assigned to each treatment.

During regular business operation, the existing scoring model assigns a score $S(x)$ to each customer, where $S(x)$ could be an estimate of the conversion probability or the ITE. The model score $S(x)$ is compared to a threshold θ to classify customers into groups, where the group *high potentials* consists of the customers with the higher score, e.g., the highest probability to respond positively to the treatment. The *high potential* group would be targeted during regular operation, while the *low potential* group would receive no treatment. Figure 6.1a visualizes a scoring model during A/B testing. For the purpose of experimental data collection, the classification and deterministic targeting is replaced by random targeting. Independent of group assignment, each customer has an equal probability to receive treatment $e(x) = e$.

The proposed process of supervised randomization (Algorithm 1) integrates the scoring model into the randomized treatment assignment. As an intermediate step, let the treatment probability be dependent on the classification by the targeting strategy as depicted in Figure 6.1b. Different to the A/B test described in Figure 6.1a, where the targeting policy does not affect



(a) Full randomization (A/B test) showing targeting policy

(b) Supervised randomization (two groups)

Figure 6.1: Experimental design of full randomization (left) and supervised randomization (right). Note the heterogeneity in treatment probability for supervised randomization. *Rand.* indicates random assignment

the treatment assignment, we now treat high potential customers with probability e_h and low potential customers with probability e_l, where $e_h \neq e_l$ and $e_l, e_h \in (0; 1)$. Note that e_h and e_l do not need to sum up to 1. We increase the treatment probability in the high potential group relative to the low potential group by choosing e_h and e_l so that $e_h > e_l$. Thereby, more high potential customers than low potential customers are treated, in accordance with the scoring model and approximating the regular targeting policy. Simultaneously, we preserve a degree of randomization in the treatment/control assignment, since each customer has some probability to be assigned to the treatment or control group, respectively. The randomization is required to fulfill the overlap assumption (Eq. 6.2) and should be large enough in practice to ensure coverage over the range of customer characteristics in both the treatment and control group. By violating the overlap assumption and setting $e_h = 1$ and $e_l = 0$, we recover the deterministic targeting policy of the classification model, where only customers in the *high potential* group are treated. Note that if we choose a constant treatment probability $e_h = e_l$, the process simplifies to an A/B test on the whole population as shown in Figure 6.1a.

We can further approximate individualized targeting by introducing more groups, each with a unique treatment probability e_k. Define a set of thresholds $[\theta_1, \theta_2, \ldots, \theta_K]$ and corresponding treatment probabilities $[e_1, e_2, \ldots, e_K]$ to target customer i with probability e_k for which $\theta_{k-1} < S(x_i) \leq \theta_k$. As above, we require $e_k \in (0; 1)$ and $\sum_k e_k = 1$. By increasing the number of thresholds K, we approximate a continuous mapping $M : S(x) \to e$, where each customer is assigned an individual treatment probability e_i proportional to the individual model score.
The specific mapping from model scores to treatment probability should follow the requirements of the application. We propose to determine the mapping by defining a set of k equal-sized intervals on the range of the model score in the training data $[\min(S(X_{train})), \max(S(X_{train}))]$ and assigning a linearly increasing treatment probability e_k to each interval, while setting the

Algorithm 1: Supervised Randomization for a Controlled Experiment with K Treatments

Input: Scoring model $S(\cdot)$; Treatment probability mapping $M(\cdot)$
Output: Treatment probability $e_{i,k}$; Treatment assignment $D_i \in \{0, 1, \ldots, K\}$; Outcome Y_i
for $i = 1, \ldots, N$ **do**
  Observe customer X_i
  Calculate customer score $s_{i,k} = S(X_i)$
  Set treatment probability $e_{i,k} = M(s_{i,k})$
  Draw treatment $D_i \sim \text{Categorical}(e_{i,k})$
  if $D_i == 0$ **then**
    Do not treat individual i
    Observe outcome $Y_i(0)$
  else
    for k *in* $1, \ldots, K$ **do**
      if $D_i == k$ **then**
        Treat individual i with treatment k
        Observe outcome $Y_i(k)$
      end
    end
  end
end

lowest treatment probability at $e_1 = 0.05$ and the highest at $e_K = 0.95$. Note that asymmetric mappings result in a controlled shift of average treatment probability. The design of the mapping thus allows the straightforward extension to imbalanced supervised randomization. We reiterate that supervised randomization (Algorithm 1) randomly assigns each customer to the treatment or control group, but adjusts the probability of this assignment based on the output of a scoring model, so that customers with higher score are treated with higher probability. The assigned individual treatment probabilities are logged and used in the subsequent analyses.

6.4.2 Inverse Probability Weighting

Using the targeting model to adjust the individual probability to receive treatment introduces a sampling bias into the experiment. The sampling bias is a direct result from the violation of independence between treatment probability and the individual characteristics via the scoring model. This type of selection bias commonly occurs in observational studies, where customers self-select into the treatment group, or in natural experiments. In both situations, the sample shows measurable distributional differences between the control and treatment group. Subsequent evaluation or model building need to correct for the selection bias to ensure unbiased estimates of the treatment effect. We will discuss IPW as a method that is easily integrated into model building and evaluation and discuss the doubly robust estimator as a recent extension. For a comprehensive overview of approaches including IPW see (Knaus et al., 2019). The idea underlying all approaches is to weight each observation in the treatment or control group by the inverse of its respective probability to be assigned to the observed group.

In contrast to observational studies where the treatment probability is estimated, the true probability at which customers receive the treatment is assigned actively based on a scoring

model and a set of observed variables and is consequently known exactly under supervised randomization. Without the need to estimate the treatment probability from the data, we avoid confoundedness due to unobserved variables or misspecification of the propensity model by design.

IPW restores the hypothetical distribution as it would look like in a fully randomized experiments by weighting every customer with regard to the individual treatment probability. Intuitively, customers who were assigned by chance to the treatment group, even though their characteristics result in a low treatment probability, are underrepresented in the treatment group. IPW assigns these customers a higher weight. For example, if the probability of being in the treatment group for a customer is $e(x) = 0.2$ then the observed outcome if this customer received treatment is multiplied by $1/e(x) = 1/0.2 = 5$. Vice versa, if the same customer was assigned to the control group, which happened with a probability of $1 - e(x)) = 0.8$, the customer's outcome in the control group is weighted by $1/0.8 = 1.25$.
The IPW corrected ATE can then be estimated as:

$$\widehat{ATE}_{IPW} = \frac{1}{N}\left(\sum_{i=1}^{N}\frac{D_i Y_i}{e(X_i)} - \sum_{i=1}^{N}\frac{(1-D_i)Y_i}{1-e(X_i)}\right). \tag{6.3}$$

In observational studies, the propensity scores are unknown and need to be estimated from observed covariates. The *doubly robust* (DR) estimator is consistent and unbiased if only one of the models, the regression or the propensity score, is correctly specified (Lunceford & Davidian, 2004):

$$\widehat{ATE}_{DR} = \frac{1}{N}\sum_{i=1}^{n}\frac{D_i Y_i - (D_i - e(X_i))\, g_1(X_i)}{e(X_i)} - \frac{1}{N}\sum_{i=1}^{n}\frac{(1-D_i)\, Y_i + (D_i - e(X_i))\, g_0(X_i)}{1-e(X_i)}. \tag{6.4}$$

Here $g_D(X_i) = \mathrm{E}(Y|D, X_i = x)$ are models of the outcome variable on x, estimated separately for $D \in \{0, 1\}$.

Adjusting for the propensity score under full randomization has no effect on the point estimate for the average treatment effect. However, there is some evidence that even in fully randomized experiments the large-sample variance of the estimate can be reduced by using estimated propensity scores to control for random imbalance in covariates as well as orthogonalization with the mean as in the doubly robust (Williamson et al., 2014).

6.5 Empirical Evaluation

We evaluate the proposed randomization procedures through a simulation study designed to represent a direct marketing setting, which is a common application of uplift modeling in marketing (see Devriendt et al., 2018; Radcliffe, 2007).[1]

[1]The `R` code for the empirical evaluation is available at https://github.com/Humboldt-WI/supervised_randomization.

Since IPW correction with the true propensity score is feasible, ATE and ITE estimates are consistent under supervised randomization. The goal of the empirical study is to compare the increased conversion rate and cost savings due to supervised randomization with the loss in efficiency due to a less balanced sample. The efficiency of each randomization procedure has two dimensions, that is, 1) monetary cost of the experiment and 2) the quality of models trained on the data collected during the experiment measured on downstream tasks.
First, the campaign profit during the experiment provides a metric on which to compare the opportunity cost of different experimental designs. We compare the campaign profit under supervised randomization to the baseline of full randomization, which provides optimal data quality, and expect opportunity costs to be lower under the proposed supervised randomization. Second, we evaluate the data generated from the experiment by comparing the predictive performance of ITE estimators trained on data under supervised randomization to the same estimators trained on data under full randomization. Our metrics of model performance are the mean absolute error to the true treatment effect (MAE), which is known in this simulation study but unknown in real-world settings, and the Qini coefficient, which is a standard metric in the uplift literature. The Qini coefficient is a rank metric similar to model lift based on the group-wise difference in conversion rates for customers ranked by their estimated treatment effect (Radcliffe, 2007).

6.5.1 Simulation Design

We compare the ATE and ITE estimates on experimental data collected under full and supervised randomization. An online evaluation of randomization procedures is challenging since it requires running a randomized experiment for each experimental design. We therefore evaluate the supervised randomization design in an offline study and leave online testing for future research. Our empirical Monte Carlo study uses real data to the extent possible to ensure a realistic setting in which we simulate the treatment effect and have full control over the treatment assignment (Knaus et al., 2019; Nie & Wager, 2017). The UCI Bank Marketing dataset (Moro et al., 2014) provides data on 45,211 customers of a Portuguese bank through 17 continuous or categorical variables covering individual socio-demographic and financial information, campaign details and macroeconomic indicators. All customers were subject to a phone marketing campaign promoting a term deposit and the target variable indicates if a customer has agreed to a deposit following the campaign.

Based on the available data, we simulate the individual treatment effect and hypothetical outcomes following the procedure of Nie and Wager (2017). The treatment effect in real data can be a complex, non-linear function of a subset of observed variables and unobserved variables (Farrell et al., 2018). Therefore, we simulate the treatment effect as a combination of the twelve variables containing personal or macroeconomic information. The treatment effect as a non-linear combination of covariates is then modelled by a neural network of one hidden layer with the number of nodes equal to the number of input variables and sigmoid activation, initialized with random weights drawn from a standard Gaussian. To simulate the existence of unobserved covariates, e.g. due to privacy concerns, we remove variables with personal information on the

customers' age and marital status from the subsequent analysis.
In marketing settings, we further expect the ATE to be positive but small and the ITE to be mostly non-negative as marketing theory suggests a direct marketing campaign to increase overall conversion, with potentially zero but rarely negative impact on customers (Hitsch & Misra, 2018). We center the simulated ITE distribution at an ATE of 5% and scale the standard deviation to 0.04 for 89% of simulated ITE to be positive. For our application, an ATE of 5% implies that the telephone campaign will convince an additional 5% of randomly targeted customer to register a term deposit. Because all customers in the observed data have received the marketing treatment, we simulate the potential outcome without treatment by flipping outcome labels for observations chosen randomly in proportion to their treatment effect as in Nie and Wager (2017).

Supervised randomization integrates an existing customer scoring model into the experimental design. A more accurate estimate from the existing model increases the extent to which potential cost savings are realized during experimentation. Noisier estimates of the scoring model lead to treatment assignment that is less profitable but has less influence on downstream tasks. In particular, supervised randomization with a noisy scoring model samples more evenly in the covariate space, thus mitigating the efficiency loss in downstream tasks, assuming a stochastic estimation error. We control the quality of the existing targeting model by simulating a noisy causal model with predictions $\hat{\tau}_i = \tau_i + \varepsilon$ and $\varepsilon \sim \mathcal{N}(0, \sigma)$. We report results for σ=0.025, such that customers with ITE equal to the ATE have a 95% chance to receive a prediction in the range [0,0.1] between the true and predicted treatment effect to provide a conservative estimate of the cost savings from supervised randomization. We split the data into four folds for cross validation and randomly assign treatment to each observation in the training data according to full or supervised randomization. For ITE estimation, we then estimate the ITE model on the training data and evaluate its prediction on the holdout fold. Since the random treatment assignment introduces additional randomness into the evaluation, we repeat the treatment assignment 50 times for each holdout fold and report the average over a total of 200 repetitions.

6.5.2 Statistical Model Performance Analysis

We first establish the effectiveness of supervised randomization independent of any application-specific cost setting. We evaluate the cost efficiency during experimentation through a comparison of conversion rates, ATE estimates based on their variance and ITE estimates based on uplift-specific performance metrics.

For cost efficiency, Table 6.2 reports the mean target fraction and conversion rate for full randomization at equal probability, full but imbalanced randomization with treatment probability 0.66 and the proposed supervised randomization procedure. We provide statistics on targeting no or all customers for context. However, targeting no or all customers and other non-randomized targeting strategies do not allow experimental data collection. In other words, settings *None* and *All* are inapplicable in practice for targeting policy evaluation or treatment effect estimation.

Table 6.2: Ratio of targeted customers and corresponding conversion rate under each randomization procedure. *None*/*All* denote targeting no/all customers for reference

	None	Full	*Supervised*	Full (Imb.)	All
Targeted Fraction of Customers	0.000	0.500	*0.500*	0.666	1.000
Conversion Rate	0.109	0.135	*0.143*	0.143	0.160

The target fractions for full randomization is 0.5 by definition and for imbalanced full randomization 0.66 by design. Small deviations from the target fraction are possible since treatment assignment is randomized on the individual level. The direction and ratio of the imbalance between the size of treatment to control group are in practice set by the experimenter to match the expected average treatment effect or marketing requirements, e.g. campaign budget. We chose a ratio of 2:1 in favor of targeting a larger group of customers following the most common design observed for customer targeting data in related research (see Table 6.1).

The conversion rate under each randomization provides an indirect measure of the campaign success with a higher conversion rate as an indicator of monetary returns. The increase in conversion rate from targeting no customers at 10.9% to targeting all customers at 16% reflects the simulated positive average treatment effect, specifically that customers are on average 5 percentage points more likely to convert after receiving the marketing treatment. During a fully randomized experiment, we observe an increase in the overall conversion rate by 2.6 percentage points to 13.5%. At the same fraction of customers targeted, the proposed supervised randomization increases the conversion rate by another 0.8 points to 14.3%. The improvement due to supervised randomization is the direct result of adjusting each customer's probability to be treated based on the targeting model and targeting customers with a high predicted treatment effect.
The benchmark strategy, imbalanced full randomization, increases treatment probability for all customers indiscriminately. The increase in individual treatment probability results in a conversion rate increase by 0.8 percentage points, identical to the increase under full randomization, but at a higher fraction of customers targeted. The managerial implication is that supervised randomization achieves the same conversion rate as current best practice, while reducing the targeting rate with its associated costs by 24%.

The higher conversion rate from targeting randomization towards relevant customers comes at the downside of collecting less data for customer groups with very high or very low treatment probability. While we can use the logged treatment probabilities to optimally correct for the sampling bias that is introduced by supervised randomization, estimates of the treatment effect will exhibit higher uncertainty through higher variance. Figure 6.2 shows the estimated ATE under each randomization procedure. We see that 1) deviations from full randomization in the form of imbalanced full randomization and supervised randomization return unbiased estimates and 2) the overall variance from the true value and the number of extreme deviations increases when moving from full randomization to supervised randomization.
A Kruskal-Wallis test verifies that there is no significant difference in the mean point estimate

among the four settings (df=3, $\chi^2 = 1.00$). We are thus able to verify the theoretical exposition and to show that the selection bias introduced by supervised randomization can be corrected for by applying either IPW or DR as described above. The additional uncertainty due to supervised randomization is less pronounced when using DR to correct for heterogeneous treatment probabilities instead of IPW. DR estimates exhibit a significantly lower variance when compared to IPW estimates, based on a Levene-test for homogeneity of variance (df=1, F=10.29).

Figure 6.2: Estimated average treatment effect averaged over 200 iterations for each randomization procedure. The dashed horizontal line denotes the simulated true average treatment effect, dots within each boxplot denote the mean estimated ATE

As uplift applications are concerned with the estimation of individualized treatment effects for customer scoring, we proceed to evaluate the model performance of two causal models on the data collected under each randomization procedure. We select the two-model approach using logistic regression and the causal forest and report the performance of using the ATE as a constant prediction for reference. Since our focus is on the comparison of the randomization procedures rather than a comparison of ITE estimators, we manually set the parameters for the causal forest as follows: number of variables tried at each split (mtry) = 7, number of trees

= 500, minimum node size = 20 and sample fraction for honest tree building = 0.5. Model predictions are evaluated using the mean absolute error to the true ITE and the Qini score on holdout data.

Table 6.3: Average profit-agnostic performance of causal models for each randomization procedure. We evaluate the models using the MAE to the (simulated) true treatment effect (lower is better) and the Qini coefficient (higher is better)

	ATE		Two-Model (LR)		Causal Forest	
	MAE	Qini	MAE	Qini	MAE	Qini
Full	0.0324	-	0.0353	0.0045	0.0276	0.0056
Full (imb.)	0.0324	-	0.0357	0.0045	0.0275	0.0057
Supervised	*0.0325*	-	*0.0383*	*0.0041*	*0.0295*	*0.0047*

We identify two takeaways in Table 6.3. First, the causal forest outperforms the two-learner approach on both MAE and Qini. The performance difference is consistent over all randomization procedures with the causal forest resulting in a MAE lower by about 0.008 points and a Qini higher by 0.001 points. Second, we observe that deviating from full randomization to supervised randomization leads to the expected decrease in model performance. Under supervised randomization, the difference to full randomization for the two-model approach is 0.003 points MAE and 0.0004 points Qini and for the causal forest 0.002 points MAE and 0.001 points Qini. Imbalanced full randomization at $e = 0.66$ shows no substantial performance decrease compared to balanced full randomization, although additional experiments indicate lower performance at higher levels of imbalance. The subsequent profit-based analysis aims to provide a comprehensible evaluation of the observed differences in a business context.

6.5.3 Profit Analysis

We proceed to empirically show the extent to which supervised randomization can reduce the cost of running a randomized experiment and the size of the expected trade-off measured by the performance of models trained on the collected data. The profit setting for telephone marketing is described by the gross profit resulting from a conversion and the variable contact cost of making a call to the customer. If we assume a constant interest margin for the bank, the gross profit from a one-year term deposit Ω is equivalent to the net interest margin m and the deposit amount A, $\Omega_i = mA_i$.
The incremental gross profit due to a marketing campaign is defined as change in the conversion probability, the treatment effect, to earn the gross profit on conversion minus the contact cost c, i.e. $\Delta\Omega_i = \tau_i mA_i - c$.
Given an accurate estimate of the treatment effect τ_i, the decision to target a specific customer is profitable when the predicted incremental gross profit for the customer is positive, i.e. $\hat{\tau}_i mA_i - c > 0$.

To simplify interpretation, we consider cost ratios in the range of $[5, 10, \ldots, 50]$ to 1. Evaluation over a range of cost settings ensures the robustness of our results and allows generalization to

a variety of profit and cost scenarios that may arise across banks or industries, e.g. for catalog marketing. We can empirically confirm the plausibility of the range of cost ratios by analyzing the ratio of customers which are targeted under each cost setting. For cost ratios below 10:1 and above 50:1, individual targeting policies are dominated by indiscriminate targeting of no or all customers, respectively. The cost ratio corresponds to different values of the interest margin m and deposit amount A at standardized contact cost. Assuming a constant amount of the term deposit $\overline{A}$ for each customer, the cost ratio can be interpreted as the ratio between the gross profit over a range of interest margins m standardized to contact costs of $c = 1$ per contact.

We evaluate the cost-saving potential of using supervised randomization during experimentation based on the campaign profit resulting from a randomized experiment for each randomization procedure. We report the campaign profit per prospective customer and the difference in campaign profit relative to full randomization in Table 6.4. As above, we include targeting no customers and targeting all customers for reference, but stress that non-randomized targeting strategies do not allow experimental data collection, making them inapplicable for causal modeling in practice.

Table 6.4: Campaign profit (per customer) for randomized experiments under each randomization procedure and across purchase margins. *None*/*All* denote targeting no/all customers for reference. *Full (Imb.)* denotes full randomization with a treatment probability of 66%

Conversion	Campaign profit per customer (€)				
Value (€)	None	Full	*Supervised*	Full (Imb.)	All
10	1.09	0.85	*0.93*	0.76	0.60
15	1.64	1.52	*1.65*	1.48	1.40
20	2.18	2.19	*2.37*	2.19	2.20
25	2.73	2.86	*3.09*	2.91	3.00
30	3.27	3.54	*3.80*	3.62	3.80
35	3.82	4.21	*4.52*	4.34	4.60
40	4.36	4.88	*5.24*	5.05	5.40
45	4.91	5.56	*5.96*	5.77	6.20
50	5.45	6.23	*6.67*	6.48	7.00

The empirical results in Table 6.4 support the proposition that supervised randomization increases the campaign profit during experimentation relative to full randomization for the full range of conversion values we consider in this study. In relative terms, supervised randomization increases the experimental campaign profit by 7.1–9.4% compared to full randomization and by 2.9–8.2% compared to imbalanced randomization.

For a conversion value of €10, we observe a marginal profit of €0.85 per customer under full randomization and a marginal profit of €0.93 under supervised randomization. The absolute increase in campaign profit is more pronounced when the cost ratio is higher. A value of €50 corresponds to a marginal profit per customer of €6.23 under full randomization compared to €6.67 under the proposed supervised randomization. Cost savings per customer compared to

full randomization amount to €0.08 and €0.44, respectively.
We translate the per customer savings to an experimental campaign of 40,000 prospective customers, who are randomly targeted. This is the size of the observed telephone marketing campaign and, with less observations than 9 of the 11 experimental marketing campaigns summarized in Table 6.1, may provide a conservative estimate. The total cost savings per experiment when replacing full randomization with supervised randomization translate to €3,200 for a marginal profit of €10, €10,400 for a marginal profit of €30 and €17,600 for a marginal profit of €50. Experiment costs and the related savings arise whenever data is collected for policy evaluation or (re-)estimation of the customer scoring model.

For conversion values greater or equal €20, targeting all customers is more profitable than not targeting any customer. The imbalanced full randomization, which we identify as standard in practice, is more profitable than full randomization only at values above €20. At these values, imbalanced randomization achieves savings of 0 to €0.25 per customer compared to full randomization for conversion values between €20 and €50, respectively. Compared to imbalanced full randomization, the proposed supervised randomization generates additional cost savings per customer of about €0.18 for all values between €20 and €50. Again translated to an experiment campaign of 40,000 prospective customers, the total cost savings per experiment of supervised randomization when compared to the industry-standard range from €7,200 for a marginal profit of €20 to €7,600 for a marginal profit of €50. Note that it is possible to combine supervised randomization with imbalanced targeting. Increasing the average treatment probability through a custom treatment probability mapping may further increase campaign profit in settings where treatment is highly profitable.

Having discussed the expected cost savings during experimentation, we next discuss the opportunity costs on downstream tasks associated with the increase in model uncertainty under supervised randomization. We first report the campaign profit per customer when customers are targeted by the two-model approach or causal forest and each model is trained on experimental data collected under the different randomization procedures.

Table 6.5 shows that the expected decrease in profit for scoring models trained on data collected under supervised randomization is small but observable in the order of 1% of the absolute campaign profit per customer. For a basket margin of €30, the two-model logistic regressions achieve a campaign profit of €3.83 per customer under full randomization and a campaign profit of €3.80 under supervised randomization, a decrease of 0.8%. The causal forest achieves a campaign profit of €3.89 when trained on data from experiments under full randomization with a decrease by 1.3% to €3.84 under supervised randomization. Compared over all values, supervised randomization induces a decrease in per customer profit between €0 and €0.04 for the two-model approach and €0 and €0.05 for the causal forest compared to full randomization.

6.6 Conclusion

Customer targeting is a continuously growing and widely studied application of scoring models. While research has focused on the prediction of future customer behavior to inform decision-

Table 6.5: Campaign profit using targeting models trained on data collected under each randomization procedure. We evaluate the campaign profit per customer over a range of cost ratios

Conversion	Two-Model (Logit)			Causal Forest		
Value (€)	Simple	Simple (Imb.)	*Supervised*	Simple	Simple (Imb.)	*Supervised*
10	1.06	1.06	*1.06*	1.09	1.09	*1.09*
15	1.65	1.65	*1.65*	1.66	1.67	*1.65*
20	2.33	2.33	*2.32*	2.36	2.36	*2.33*
25	3.07	3.06	*3.05*	3.11	3.12	*3.08*
30	3.83	3.82	*3.80*	3.89	3.89	*3.84*
35	4.60	4.60	*4.57*	4.67	4.67	*4.62*
40	5.38	5.38	*5.35*	5.45	5.45	*5.41*
45	6.16	6.16	*6.13*	6.24	6.24	*6.20*
50	6.95	6.95	*6.91*	7.03	7.03	*7.00*

making, a growing research stream has established uplift models to estimate the causal effect of a marketing action on each customer based on observed customer characteristics. The training and evaluation of causal models require data collected through experiments, in which customers are randomly assigned to treatments. However, experimental data collection incurs high costs by temporarily replacing an established targeting policy with random targeting.

We propose supervised randomization as a solution to reduce the cost of experimentation by integrating an existing scoring model into the experimental design. By mapping model scores to individual treatment propensities, we are able to target more profitable customers while maintaining stochastic treatment assignment. An empirical Monte Carlo study on telemarketing shows that supervised targeting can reduce the cost of an experimental campaign on 40,000 prospective customers by 7.1–9.4% compared to full randomization and 2.9–8.2% compared to imbalanced randomization, depending on the specific profit-cost ratio.

Active management of treatment assignment during experimentation leads to an overrepresentation of profitable customers in the treatment group, which causes selection bias when standard estimators are applied to estimate treatment effects. We consequently summarize inverse probability weighting and doubly robust estimation as well-studied methods to correct for selection bias when estimating average and individualized treatment effects. We show that the estimated treatment effects are unbiased and provide indicators of the increase in uncertainty related to supervised randomization. Empirical evaluation indicates that higher uncertainty of the scoring model may lead to a decrease in campaign profit by 0.8–1.3% depending on the specific profit-cost ratio. Further evaluation in real-world experiments is necessary to establish net cost savings in practice.

Overall, we argue that the methodology developed in the medical and econometric literature has not yet been fully studied and applied in the uplift setting. Doubly robust estimation serves as one example of a wider set of tools to correct for selection issues in the data. We

further identify experimental data collection as a fundamental part of causal modeling. We expect that supervised randomization provides a first step towards a wider analysis of practical experimental design.

Bibliography

Ascarza, E. (2018). Retention futility: Targeting high risk customers might be ineffective. *Journal of Marketing Research*, *55*(1). https://doi.org/10.1509/jmr.16.0163

Ascarza, E., Ebbes, P., Netzer, O., & Danielson, M. (2017). Beyond the target customer: Social effects of customer relationship management campaigns. *Journal of Marketing Research*, *54*(3), 347–363. https://doi.org/10.1509/jmr.15.0442

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360. https://doi.org/10.1073/pnas.1510489113

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178.

Athey, S., & Wager, S. (2017). Efficient Policy Learning. *arXiv preprint*, arXiv:1702.02896.

Chickering, M., & Heckerman, D. (2000). A Decision Theoretic Approach to Targeted Advertising, In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI'00)*, Morgan Kaufmann.

Devriendt, F., Moldovan, D., & Verbeke, W. (2018). A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. *Big Data*, *6*(1), 13–41. https://doi.org/10.1089/big.2017.0104

Diemert, E., Betlei, A., Renaudin, C., & Amini, M.-R. (2018). A Large Scale Benchmark for Uplift Modeling, In *Proceedings of the AdKDD and TargetAd Workshop, KDD*, London, United Kingdom, ACM.

Farrell, M. H., Liang, T., & Misra, S. (2018). Deep Neural Networks for Estimation and Inference: Application to Causal Effects and Other Semiparametric Estimands. *arXiv e-prints*, arXiv:1809.09953.

Gordon, B. R., Zettelmeyer, F., Bhargava, N., & Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Science*, *38*(2), 193–364. https://doi.org/10.1287/mksc.2018.1135

Gubela, R. M., Bequé, A., Gebert, F., & Lessmann, S. (2019). Conversion uplift in e-commerce: A systematic benchmark of modeling strategies. *International Journal of Information Technology & Decision Making*, *18*(3), 747–791.

Gubela, R. M., Lessmann, S., Haupt, J., Baumann, A., Radmer, T., & Gebert, F. (2017). Revenue Uplift Modeling, In *Proceedings of the 38th International Conference on Information Systems (ICIS)*, AIS.

Guelman, L. (2014). *Optimal Personalized Treatment Learning Models with Insurance Applications* (Doctoral Thesis). Universitat de Barcelona.

Guelman, L., Guillén, M., & Pérez-Marín, A. M. (2015). Uplift Random Forests. *Cybernetics and Systems*, *46*(3-4), 230–248. https://doi.org/10.1080/01969722.2015.1012892

Hansotia, B. J., & Rukstales, B. (2002). Direct marketing for multichannel retailers: Issues, challenges and solutions. *Journal of Database Marketing & Customer Strategy Management*, *9*(3), 259–266. https://doi.org/10.1057/palgrave.jdm.3240007

Hansotia, B., & Rukstales, B. (2002). Incremental value modeling. *Journal of Interactive Marketing*, *16*(3), 35–46. https://doi.org/10.1002/dir.10035

Hillstrom, K. (2008). *The MineThatData E-Mail Analytics and Data Mining Challenge.*

Hitsch, G. J., & Misra, S. (2018). Heterogeneous Treatment Effects and Optimal Targeting Policy Evaluation. *SSRN*.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 663–685.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5–86. https://doi.org/10.1257/jel.47.1.5

Kane, K., Lo, V. S. Y., & Zheng, J. (2014). Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. *Journal of Marketing Analytics*, *2*(4), 218–238. https://doi.org/10.1057/jma.2014.18

Knaus, M. C., Lechner, M., & Strittmatter, A. (2019). Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence. *IZA Discussion Paper*, *12039*.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165.

Lachin, J. M., Matts, J. P., & Wei, L. J. (1988). Randomization in clinical trials: Conclusions and recommendations. *Controlled Clinical Trials*, *9*(4), 365–374. https://doi.org/10.1016/0197-2456(88)90049-9

Lo, V. S. Y. (2002). The true lift model: A novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter*, *4*(2), 78–86.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, *23*(19), 2937–2960.

Marco Caliendo, Michel Clement, Dominik Papies, & Sabine Scheel-Kopeinig. (2012). The cost impact of spam filters: Measuring the effect of information system technologies in organizations. *Information Systems Research*, *23*(3), 1068–1080.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31. https://doi.org/10.1016/j.dss.2014.03.001

Nie, X., & Wager, S. (2017). Quasi-Oracle Estimation of Heterogeneous Treatment Effects. *arXiv e-prints*, 1712.04912.

Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems*, *52*(2), 464–473. https://doi.org/10.1016/j.dss.2011.10.007

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., & Tibshirani, R. (2018). Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine*, *37*(11), 1767–1787. https://doi.org/10.1002/sim.7623

Radcliffe, N. J. (2007). Using control groups to target on predicted lift: Building and assessing uplift models. *Direct Marketing Analytics Journal*, 14–21.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*(427), 846–866.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rosenberger, W. F., & Lachin, J. M. (1993). The use of response-adaptive designs in clinical trials. *Controlled Clinical Trials*, *14*(6), 471–484. https://doi.org/10.1016/0197-2456(93)90028-C

Rzepakowski, P., & Jaroszewicz, S. (2012). Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, *32*(2), 303–327. https://doi.org/10.1007/s10115-011-0434-0

Schnabel, T., Swaminathan, A., Singh, A., Chandak, N., & Joachims, T. (2016). Recommendations as Treatments: Debiasing Learning and Evaluation, In *Proceedings of the 33rd International Conference on Machine Learning*.

Schulz, K. F., & Grimes, D. A. (2002). Generation of allocation sequences in randomised trials: Chance, not choice. *The Lancet*, *359*(9305), 515–519. https://doi.org/10.1016/S0140-6736(02)07683-3

Schwartz, E. M., Bradlow, E. T., & Fader, P. S. (2017). Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, *36*(4), 500–522. https://doi.org/10.1287/mksc.2016.1023

Statista. (2017). *Advertising Spending in the Catalog, Mail-order Houses Industry in the United States* (tech. rep.).

Statista. (2019). *eCommerce* (tech. rep.).

Swaminathan, A., & Joachims, T. (2015). Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, *16*, 1731–1755.

Williamson, E. J., Forbes, A., & White, I. R. (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine*, *33*(5), 721–737.

Zaniewicz, L., & Jaroszewicz, S. (2013). Support Vector Machines for Uplift Modeling, In *13th International Conference on Data Mining Workshops*, IEEE.

Chapter 7

The Price of Privacy: An Evaluation of the Economic Value of Collecting Clickstream Data

Publication

Baumann, A., Haupt, J., Gebert, F., & Lessmann, S. (2019). The price of privacy: An evaluation of the economic value of collecting clickstream data. Business & Information Systems Engineering, 61(4), 413–431. https://doi.org/10.1007/s12599-018-0528-2

Abstract

The analysis of clickstream data facilitates the understanding and prediction of customer behavior in e-commerce. Companies can leverage such data to increase revenue. For customers and website users, on the other hand, the collection of behavioral data entails privacy invasion. The objective of the paper is to shed light on the trade-off between privacy and the business value of customer information. To that end, we review approaches to convert clickstream data into behavioral traits, which we call clickstream features, and propose a categorization of these features according to the potential threat they pose to user privacy. We then examine the extent to which different categories of clickstream features facilitate predictions of online user shopping patterns and approximate the marginal utility of using more privacy adverse information in behavioral prediction models. This way, the paper links the literature on user privacy to that on e-commerce analytics and takes a step toward an economic analysis of privacy costs and benefits. In particular, the results of empirical experimentation with large real-world e-commerce data suggest that the inclusion of short-term customer behavior based on session-related information leads to large gains in predictive accuracy and business performance, while storing and aggregating usage behavior over longer horizons has comparably less value.

Chapter 8

E-Mail Tracking: Status Quo and Novel Countermeasures

Publication
Bender, B., Fabian, B., Lessmann, S., & Haupt, J. (2016). E-Mail Tracking: Status Quo and Novel Countermeasures. Proceedings of the 37th International Conference on Information Systems (ICIS).

Abstract
E-mail advertisement, as one instrument in the marketing mix, allows companies to collect fine-grained behavioural data about individual users' e-mail reading habits realised through sophisticated tracking mechanisms. Such tracking can be harmful for user privacy and security. This problem is especially severe since e-mail tracking techniques gather data without user consent. Striving to increase privacy and security in e-mail communication, the paper makes three contributions. First, a large database of newsletter e-mails is developed. This data facilitates investigating the prevalence of e-mail tracking among 300 global enterprises from Germany, the United Kingdom and the United States. Second, countermeasures are developed for automatically identifying and blocking e-mail tracking mechanisms without impeding the user experience. The approach consists of identifying important tracking descriptors and creating a neural network-based detection model. Last, the effectiveness of the proposed approach is established by means of empirical experimentation. The results suggest a classification accuracy of 99.99%.

8.1 Introduction

Data about e-mail reading behaviour can be used to infer valuable commercial information. In marketing, for example, it allows user preferences to be derived and the reach and effectiveness of e-mail marketing campaigns to be measured (Hasouneh & Alqeed, 2010). Contemporary e-mail tracking techniques enable the sender to track how often an e-mail is read, which device the recipient uses, and the time as well as location from which the e-mail is read (Fabian et al., 2015). Importantly, this information is typically gathered without the recipient's consent or acknowledgement. Tracking can also constitute a security threat. Spammers and hackers commonly rely on e-mail tracking to detect and collect active e-mail addresses for their illegal activities. From an end-user perspective, e-mail tracking procedures therefore involve various security and privacy issues.

Mail users should be equipped with effective and reliable protection methods. In order to provide advancements towards user privacy and security protection, this paper follows the design science research paradigm (Peffers et al., 2007). The study starts with a survey of relevant literature and proposes a definition for e-mail tracking. Following that, e-mail tracking technology is explained. Further contributions involve the experimental analysis of information that can be gathered using e-mail tracking, and a critical comparison of currently available protection measures. Then, with regard to problem identification and relevance, our paper presents a large empirical study, confirming e-mail tracking as an important and widespread privacy issue.

This motivates another major contribution of our research: the design of countermeasures, encompassing the development of a novel method for tracking-image identification that is based on machine learning. A demonstration and evaluation are realised through a quantitative and empirical evaluation of the developed detection model based on a large dataset of over 4,500 mails from 300 global companies, including more than 110,000 images. This article will serve to communicate our results.

8.2 Definition and Related Work

E-mail tracking and its impact on privacy are often mentioned in the general press, for example in conjunction with scandals that have been uncovered using e-mail tracking technologies (Evers, 2006). As far as the academic literature is concerned, surprisingly few papers have looked into the topic (Bonfrer & Drèze, 2009; Fabian et al., 2015; Hasouneh & Alqeed, 2010) and these do not focus on countermeasures for e-mail tracking. Some initiatives to develop anti-tracking software have been undertaken in corporate practice. However, as we show below, they do not provide sufficient protection. The lack of effective countermeasures motivates this research, which emphasises the technical and process-related aspects of e-mail tracking. In accordance with this focus and with inspiration from Fabian et al. (2015), we propose the following definition of the term e-mail tracking: *E-mail tracking allows mail senders to gather information on an individual recipient's reading behaviour of single mails without the need for any further interaction or the recipient's permission.*

Some characteristics of the definition deserve further clarification. *Individual recipient*: Relevant techniques allow the gathering of information on the individual recipient's behaviour. This is important in order to distinguish e-mail tracking procedures from general aggregated traffic measuring techniques. *Reading behaviour*: The minimum requirement is that a technique provides information about whether a single mail has been opened by a specific recipient. *Single mails*: To fulfil the requirement of marketers or other trackers, a tracking mechanism provides information on the level of single mails. In combination with the *individual recipient* requirement, this allows trackers to infer the reading behaviour of every recipient for every mail that was issued. *Without any further interaction*: This emphasises methods that do not require further user actions than simply opening an e-mail. One technical implication of this understanding is that we concentrate on tracking pixels but not on tracking links which users need to click on (Fabian et al., 2015). The important aspect is that simply opening the mail is suffi-

cient to trigger the tracking mechanism. *Without recipient's permission*: This emphasises the fact that the mechanism does not require any acknowledgement of the recipient; the technique therefore distinguishes itself from functions such as mail return receipts. This characteristic involves possibilities of secret surveillance.

From a technological point of view, e-mail tracking can be understood as an adaptation of web tracking mechanisms to HTML-based e-mails. Unlike e-mail tracking, web-tracking mechanisms have received much attention in the literature. The *use* of web tracking in different situations (Javed, 2013; Jensen et al., 2007) as well as their *detection* (Alsaid & Martin, 2003; Fonseca et al., 2005) have been analysed. *Prevention* of such mechanisms, including evaluation of software solutions, has also been the topic of studies (Fonseca et al., 2005; Leon et al., 2012). Other research emphasises the technical aspects of web tracking, such as different categories of web bugs (Dobias, 2011) or the potential for aggregating multiple server log files (Evans & Furnell, 2003). Yet another stream of research aims at supporting website operators through reviewing and developing criteria for web tracking software selection (Fourie & Bothma, 2007; Nakatani & Chuang, 2011) or, more generally, evaluating the market for web tracking software (Krishnamurthy & Wills, 2009).

Some web tracking papers hint at the possibility of applying tracking techniques to HTML-based e-mails (Bouguettaya & Eltoweissy, 2003; Harding et al., 2001; Martin et al., 2003; Moscato et al., 2013). However, none of these studies provide further details of such an undertaking or discusses the peculiarities of e-mail tracking. Clearly, some similarity between web and e-mail tracking mechanisms exists, especially in terms of tracking technology. From an organisational point of view, a similarity may also be seen in the fact that the tracking infrastructure can be operated by the company or a contracted service provider, though in-house solutions seem to be extremely scarce in web tracking contexts (Burkell & Fortier, 2013; Sipior et al., 2011; Waisberg & Kaushik, 2009). However, one important difference concerns tracking precision. With e-mail tracking, any information can be easily linked to the user's e-mail address, which is an almost unique identifier of the user. Consequently, tracking users across devices, locations, channels etc. is much easier compared to web tracking, which heavily depends on the browser environment and its configuration. In this sense, e-mail tracking can be considered even more privacy intrusive, which further supports the need for effective countermeasures.

8.3 E-Mail Tracking Technology

To give an overview of e-mail tracking methodology and the degree to which it impedes user privacy, the following sections review the e-mail tracking process and detail how and which information is captured about mail recipients. This provides a foundation for developing effective countermeasures.

8.3.1 E-Mail Tracking Process

The tracking process (Figure 8.1) is based on e-mails that reference external resources. Therefore, it starts with the preparation of an HTML-based e-mail by the sender, since plain-text

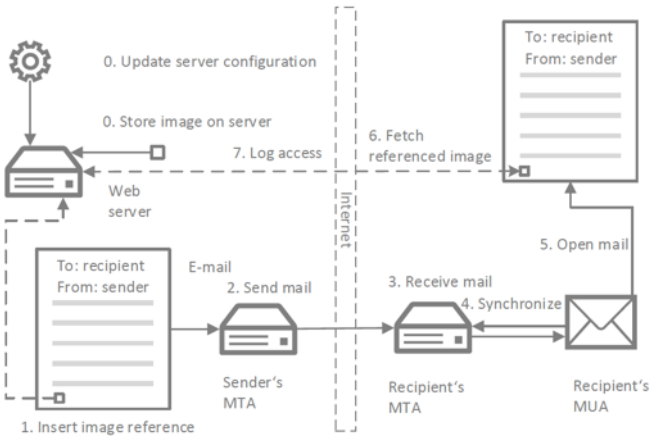


Figure 8.1: E-mail tracking operation mode

e-mails do not facilitate such references. This e-mail, which includes a tracking-image reference, passes several mail transfer agents (MTAs) until reaching the receiver's MTA. Next, the recipient opens a mail client, which synchronises the local mail repository with the newest version of the recipient's MTA. When the recipient opens the e-mail with a tracking image, the mail client requests the image from the referenced destination. The web server logs this request and provides the image to the recipient's client. Afterwards, log analysis allows information to be deducted on the recipient's e-mail reading behaviour. For example, if the e-mail is opened on different devices, every individual access is logged, which allows for cross-device tracking.

Even though the structure of tracking image references varies, the following anonymised reference serves as an example of tracking image references: http://www.example.com/action/view/3827/rtg2ryw3.

8.3.2 Information Gathered by E-Mail Tracking

To elaborate on the collection of behavioural data via e-mail tracking, we distinguish between primary and secondary information. The former is directly extracted from web server access logs, whereas the latter can be derived from combining primary information with auxiliary data sources.

In order to assess the extent of primary information available to a tracker, we constructed a prototypical tracking environment, which includes an Apache webserver to log data relevant to e-mail tracking. The entries in the server log file provide seven major pieces of information: (1) the Internet Protocol (IP) address of the host that requests the image file, (2) the date and time of the file request, (3) the request itself, which includes the URL and GET variables, (4) the status code of the request, (5) the amount of bytes that have been sent in response, (6) the referrer URL from the client, and (7) a string characterising the user agent. Furthermore, when a file is requested multiple times (i.e., it generates multiple entries), it allows information to be derived with respect to a user's reading behaviour. In our test environment, a new log entry was created every time an e-mail was opened.

With respect to secondary information, the first possibility is to induce the fact that the user

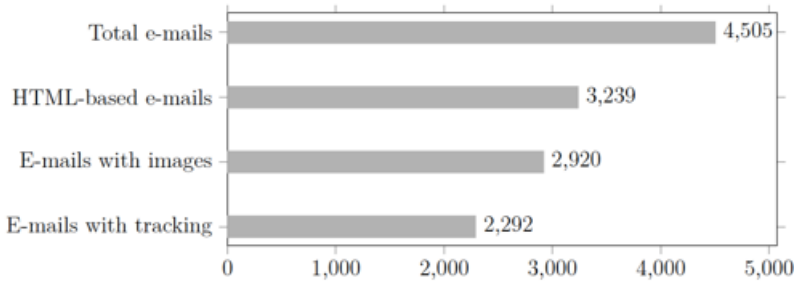


Figure 8.2: Dataset overview: tracking/non-tracking

read or at least opened the e-mail. As we show in more detail below, this follows from the fact that current e-mail clients do not download images before the corresponding e-mail is opened. Accordingly, the existence of multiple log entries allows for the conclusion that the e-mail has been opened multiple times. The combination of multiple entries for one mail, as well as multiple entries from one user for different mails, provides insight into the recipient's e-mail reading behaviour.

Furthermore, possibilities exist for identifying whether an e-mail has been forwarded. The usage of IP geolocation in combination with a log entry aggregation allows the detection of forwarded mails. HP, for example, used this technique to investigate the release of confidential information (Evers, 2006). Special log entries allow one to determine whether an e-mail has been printed (Campaign Monitor, 2010). It is also possible to gather information about the user environment by analysing the user agent string, which is part of a log entry (Agosti & Di Nunzio, 2007). Location-related information can be gathered using geolocation services (Poese et al., 2011). Based on a reverse lookup of an IP address, a log entry may also help determine a user's affiliation to a company or institution. These examples illustrate only some options for trackers and the potential for gaining insights into user behaviour through combining and correlating tracking data with external information.

8.4 International Study on E-Mail Tracking Usage

Having established the intrusiveness of e-mail tracking, a relevant follow-up question concerns the prevalence of tracking mechanisms. To answer this, we collected a unique empirical dataset of e-mail newsletters, which allowed us to evaluate the status quo of e-mail tracking. Although potentially not representative of companies' marketing communication in general, newsletter e-mails are a suitable vehicle for this analysis. First, the wide availability of different newsletters simplifies systematic data collection and facilitates the gathering of a large amount of data. Second, it seems likely that companies use e-mail tracking to assess the effectiveness of their newsletters (Hasouneh & Alqeed, 2010). To further increase this likelihood, we concentrate on larger companies because these are on average faster to adopt novel technology (Premkumar & Roberts, 1999). Contrary to Fabian et al. (2015) who performed a comparable analysis among 64 German companies, we adopted an international scope and gathered e-mail newsletters from the top-100 companies (ranked by revenue) in Germany (GER), Great Britain (GB), and the United States (USA).

Table 8.1: Tracking elements per country

Country	Tracking elements				Total
	0	1	2	3	
GER	231	1206	5		1442
UK	1173	300	16		1489
US	665	636	107	20	1428
Various	144	2			146
Total	2213	2144	128	20	4505

To gather the data, we create two identities and corresponding e-mail addresses using Gmail. With each account, we signed up for the newsletters of the pre-selected 300 companies and collected e-mails in a 13-week period (calendar week 22–34) in 2015. To identify tracking elements, we compared the e-mails received on each account. That is, we examined the HTML content of each pair of e-mails sent to matched accounts and searched for deviations in image URLs. To obtain ground truth data, we classified images for which the referral URLs do not match as tracking image, and all others as non-tracking. Last, we manually checked all tracking elements to avoid classification errors. However, to avoid bias from senders changing their e-mail policy in response to the reading behaviour of users they were tracking, we ensured that none of the external images were actually requested from the web server.

In total, each artificial identity received 4,505 e-mails, 1,442 (32%) of which came from Germany, 1,489 (33%) from the United Kingdom (UK), and 1,428 (32%) from the United States (US). The remainder, referred to as *various* below, consisted of e-mails sent from multiple countries. This usually applies if externally contracted third parties send mails for several clients in different countries from the same address.

The fraction of HTML compared to plain text e-mails is interesting since only HTML-based e-mails facilitate tracking. Out of 4,505 e-mails, 1,266 (28%) were in plain-text format, while the remaining 3,239 (72%) were HTML-based. Considering the HTML e-mails, 2,920 (90%) contained external image references. These e-mails could facilitate tracking. The HTML e-mails contained references to 110,080 external images, with an average of 38 external images per e-mail. 18% of the e-mails contained a single external image. Figure 8.2 gives an overview of the key measures relevant for e-mail tracking: 2,292 e-mails contained tracking elements, which equated to a ratio of 51% (71%) among all e-mails (HTML e-mails).

Our results reveal that the tracking quote and the fraction of HTML-based e-mails varied across countries (Table 8.1). The proportion of HTML e-mails was 88% (1,375), 34% (513), and 84% (1,205) for Germany, the UK, and the US, respectively. Concentrating on these e-mails, the tracking quote was the highest in Germany (88%), and rather similar for the US (63%) and the UK (62%).

The main options for performing e-mail tracking are in-house systems and contracted service providers, the latter of which prevail in web tracking. To shed light on the frequency of the two options in e-mail tracking, we differentiated between internal and external tracking. We defined

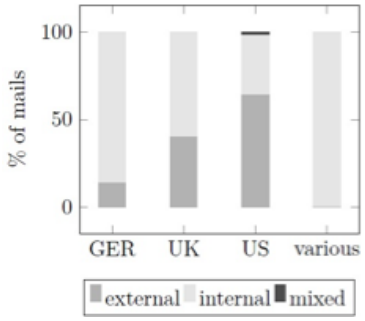


Figure 8.3: Tracking type distribution per country

Table 8.2: Tracking type distribution per country

	Tracking type			
Country	External	Internal	Mixed	Total
GER	166	1045		1211
UK	127	189		316
US	485	261	17	763
Various		2		2
Total	778	1497	17	2292

tracking as internal if the web server hosting the tracking image belonged to the company that sent the newsletter, and as external otherwise. To handle ambiguous cases, we defined a third category, 'various', which subsumed e-mails with multiple tracking images from internal and external web servers. Figure 8.3 and Table 8.2 show the results of this analysis. It reveals that Germany had the highest internal tracking rate at 86%, followed by the UK with 60%, and the US with 34%. Only mails from the US used internal and external tracking at the same time.

We used geolocation services to assess the tracking-server location. Table 8.3 shows the results for all tracking elements and corresponding web servers. Specifically, it depicts the two-letter ISO3166-1:2013 abbreviation of the country and the number of occurrence for main servers, and tracking servers in each country. As noted above, references to external images in a single e-mail may point to multiple servers. We defined the main server as the one that hosts the

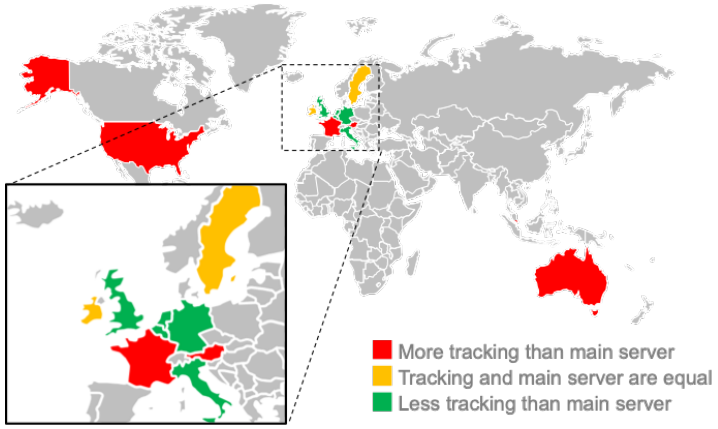


Figure 8.4: Country classification

Table 8.3: Image server locations

Country	ISO 3166	Tracking	Main Content
Austria	AT	77	52
Australia	AU	6	0
Belgium	BE	0	1
Germany	DE	913	1399
France	FR	39	1
Great Britain	GB	68	143
Ireland	IE	14	14
Italy	IT	0	23
Netherlands	NL	9	107
Sweden	SE	1	1
Singapore	SG	2	0
United States	US	1331	719

majority of external image references. According to our data, tracking images are rarely hosted on the main server if multiple servers occur in a single e-mail. For example, this situation may arise when a company stores images on an internal server but has outsourced tracking to an external service provider. To clarify the frequency of such an approach, Table 3 distinguishes between main and tracking servers.

Table 8.3 suggests that most countries differ in the number of occurrences for tracking and main server locations. One possible explanation is the use of external e-mail tracking providers that operate their business in a different country. Another reason might involve different regulations with regard to tracking technologies. In terms of server locations, Figure 8.4 classifies countries into three groups according to whether they contain more, the same, or fewer tracking servers than main servers.

Countries that host more main servers than tracking servers are Germany, Great Britain, Italy, and the Netherlands. We aim to investigate in future work whether regulations related to tracking techniques are stricter in these countries, or if there are other reasons that tracking services are hosted abroad. In Italy and Belgium, only main images but no tracking images are hosted. Another group of countries host more tracking servers than main servers, such as Austria, Australia, France, and the United States. Singapore and Australia are especially interesting since only tracking images but no main images are hosted in these countries according to our dataset. Foreign tracking servers are least common for e-mails from US companies, where 99.9% of the tracking images are hosted within the nation.

8.5 Countermeasure Conceptualisation and Review

The previous analysis shows that e-mail tracking is a common phenomenon, which emphasises the importance of countermeasures to increase privacy and security of individual users. In the following, we conceptualise technological options for countermeasure design. We also review existing implementations and discuss their merits and limitations in order to verify the necessity

of developing a novel approach.

8.5.1 Classification of Countermeasures

To summarise the sphere of possible solutions, we distinguished between deceptive and preventive countermeasures. We further divided the latter into holistic and selective approaches. Deceptive countermeasures strive to hide or modify the information sent to a mail sender or a third-party tracking provider so that these obtain selected, modified, or deliberately corrupted information. As Figure 8.1 shows, this strategy can be implemented through introducing a proxy server in the communication between e-mail sender and receiver, which caches the referenced images. The role of the proxy is to download and cache all images referenced in an in-coming e-mail prior to transferring it to the recipient's mail client. The sender of a tracking e-mail will then recognise the first access of a tracking image through the proxy, but will not be able to observe subsequent requests due to multiple opens. More importantly, the tracked identity is that of the proxy, whereas the reading behaviour of actual recipients remains unobservable. Figure 8.5 depicts two slightly different versions of this approach. In the first version (Figure 8.5a), the proxy modifies the incoming mail so that the formerly externally referenced images are included in the mail. In the second version (Figure 8.5b), the proxy server caches the externally referenced images and the references are changed so that they point to the proxy server. Every time the mail is opened, the images are fetched from the proxy and not the original tracking server.

(a) Including images into e-mail

(b) Caching images on external server

Figure 8.5: Deceptive prevention approach

The strength of the proxy-based approach is the possibility to conceal some tracking information. Weaknesses are related to the fact that the referenced images, including tracking images, are still accessed. The sender is therefore able to gather some, though potentially fuzzy, information. Another weakness is the necessity of server-side support. Operating a proxy server appears prohibitively expensive for an individual user. Even among e-mail clients, we were unable to identify an actual implementation of the proxy-based approach and, more generally, any deceptive countermeasure specific to e-mail tracking. Even though Google Mail introduced proxy technology for their web client in 2013 and hides some information such as the IP address from trackers, we conducted experiments that showed that every mail open is still registered by the tracking software. Therefore, proxy-based deception approaches do not yet provide full tracking protection.

A second option for countering tracking is to block all external content referenced in an e-mail. We call this approach holistic prevention (Figure 8.6). Given that all tracking procedures known

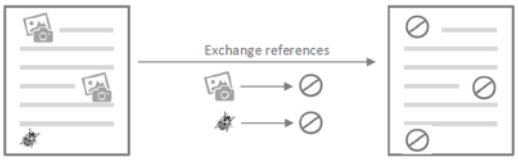


Figure 8.6: Holistic prevention approach

Figure 8.7: Elective prevention approach

today depend on references to external content, the strength of the block-all approach is a fully-reliable protection against tracking. From a technical point of view, ease of implementation on either the server's or the client's side may be considered as another advantage. In fact, most contemporary mail clients allow users to block external content. However, the massive loss of information and decrease in user experience resulting from the exclusion of all referenced images and the corresponding content constitute severe disadvantages. For example, consequences of the holistic prevention approach include incorrect formatting, loss of styling elements, and potentially misinterpretation if external images convey important information and/or are crucial to correctly interpret an e-mail message.

Finally, selective countermeasures are based on identifying and blocking tracking elements within an e-mail (Figure 8.6). The idea is to categorize the referenced images into content-providing and tracking images. Assuming that images of the former category do not provide tracking functionality, they remain untouched, whereas tracking images are removed from the e-mail. The strength of the identify-and-block concept is the combination of systematic prevention of tracking images while preserving the full user-experience. The possible pitfall is the risk of misidentification. The solution will work only as well as the algorithm for tracking-image identification.

Based on the review of alternative options for countermeasure design, we argue that a selective identify-and-block strategy provides the best balance between preventing user tracking and sustaining user experience. Accordingly, we concentrate on this approach in the remainder of the paper. In the next section, we report the results of an empirical evaluation of available implementations of the identify-and-block concept. Given the criticality of distinguishing content-providing and tracking images with high accuracy in this approach, we will then present a prototype of an identification engine and an empirically test of its effectiveness.

8.5.2 Selective Prevention – Empirical Experiments

Some mail clients and add-ons to mail clients support selective prevention. To examine the appropriateness of existing implementations, we created a testing infrastructure including e-mail accounts, e-mail clients, operating systems, hardware, and a third-party analysis tool

Table 8.4: E-mail clients usage share and protection against tracking images

Category	Mail Client	Share	Before Mail Open	After Mail Open	After Display Images	Sufficient Protection
Desktop	Outlook	8%	✓	✓	✗	✗
	Apple Mail	8%	✓	✗	–	✗
	Windows Live	2%	✓	✓	✗	✗
Mobile	iPhone Mail	28%	✓	✗	–	✗
	iPad Mail	11%	✓	✗	–	✗
	Android Mail	8%	✓	✓	✗	✗
Webmail	Gmail	18%	✓	✗	–	✗
	Outlook.com	5%	✓	✓	✗	✗
	Yahoo!	4%	✓	✓	✗	✗

(https://emailprivacytester.com/). The latter sent e-mails with tracking elements to our accounts and reported the information gathered through tracking. Using this infrastructure, we assessed the reactions of nine popular mail clients. More specifically, the selection of clients drew inspiration from a study by Litmus Labs and encompassed desktop, mobile, and web clients. We selected the three most popular mail clients in each category for analysis. Together, these clients were responsible for 92% of all mail openings (Litmus Labs, 2015). It is worth noting that the mobile category had the greatest share with 47% of mail openings, which highlights the importance of mobile devices in today's e-mail communication.

Table 8.4 reports the results of our mail client evaluation, which distinguishes three process steps: before mail open, after mail open, and after displaying images. The first step characterises the point at which an e-mail client has synchronised with the mail server and the mail is visible in the list of incoming mails but has not been opened. After mail open is the moment when the mail is selected and opened but no further action has been performed. After display images refers to the time when the e-mail has been opened and the display of external images has been allowed by the user. This is only applicable if the mail client does not download external images by default. A check mark means that the mail client does not fetch the tracking image and therefore provides sufficient protection against tracking. In contrast, a cross indicates that the tracking images have been downloaded and tracking data can be gathered.

Table 8.4 reveals that no mail client fetched referenced images prior to opening an e-mail. Afterwards, however, four out of nine clients fetched referenced images directly. The other five mail clients fetched the tracking image in the third step. Given that none of the mail clients filtered tracking elements, we concluded that they failed to provide sufficient tracking protection.

With respect to add-ons, the second approach in the selective prevention category, we identified two representatives called *Uglyemail* and *Pixelblock*. The Uglyemail browser add-on allows the detection of tracking pixels from nine third-party tracking providers, but does not offer functionality to selectively block these images. Pixelblock blocks images with a one-pixel size, but fails

to identify tracking images of different size. Both add-ons share the additional disadvantage that usage is restricted to the Google Chrome browser and Google's mail service Gmail. Given their restricted applicability, we conclude that browser add-ons also fail to provide appropriate tracking protection.

In summary, none of the tracking prevention solutions provided sufficient, practical and universal protection against e-mail tracking, which confirms the demand for novel solutions. Therefore, we proceeded with designing a prototype for a selective countermeasure and the corresponding identification engine in particular.

8.6 Tracking Image Detection

Tracking image detection can be interpreted as the classification of unknown images. All images that are referenced within an e-mail are used as input, and the classification process needs to decide whether or not each individual image is a tracking image. The detection process can be conceptualised in two steps. First, the identification of essential characteristics that distinguish normal from tracking images will result in a detection model that is based on various image attributes. Second, a classification decision needs to be made based on the detection model. This two-step approach ensures a certain independence of the detection model from the decision model. As a result, multiple different detection and decision techniques can be used and easily compared.

The detection model proposed in this paper is based on six different categories of data. The first three categories (image attributes, reference structure, and e-mail structure) subsume aspects that are directly associated with the source code of an e-mail which is part of the e-mail body. The fourth category (image server) is associated with the servers that host the images. The fifth category uses information from the whole dataset to assess whether a server is a tracking server. The sixth category covers the e-mail header.

In order to gather information on tracking images and to allow for a later validation of the developed model, the total dataset was divided into a training and a test set. The training set was used to gather information on tracking pixels and to train the detection model. The training set accounted for 76% of the total mails, while the other 24% were assigned to the test set, which was used for validation purposes.

8.6.1 Image Attributes

The first category covers image attributes that are directly associated with an image element as well as attributes referring to centrally defined style information from Cascading Style Sheets (CSS). Typically, images are embedded using the <img> tag of the Hypertext Markup Language (HTML) (Musciano & Kennedy, 2006). The analysis and model conception were restricted to attributes occurring in at least 1% of the images, since it would be difficult to derive representative results from a smaller amount of cases. Furthermore, the more attributes there were to be analysed, the slower the performance of the whole solution.

The *border* attribute defines the thickness of the border that is drawn around an image (Musciano & Kennedy, 2006). While the border attribute only occurred in one third of all tracking images, it occurs in more than three-quarters of all non-tracking images. In our dataset, all values different from zero (or an empty value) occurred in non-tracking images only. The border attribute could therefore be used to identify images clearly as non-tracking images if their border was provided and deviated from zero.

The *width* attribute allows the horizontal size of a displayed image to be specified once the website or e-mail has been rendered (Musciano & Kennedy, 2006). 65% of the tracking pixels with a specified width had a width of one. The *height*, similar to the width, allows the height to be defined at which an image is displayed. Another result of our empirical analysis is that the vast majority of tracking pixels are quadratic. One assumption from related work was that tracking pixels have an area less than 10 and are usually very small (Fabian et al., 2015). However, our dataset also contains 27 tracking pixels with a specified area of more than 10, which therefore does not match the former model criteria.

The *style properties* that were considered in the analysis are composed of the style attribute tag and centrally provided CSS commands. In order to avoid an individual discussion of each attribute and to be able to dynamically expand the attribute classification, categories of CSS commands were set up, subsuming commands which fulfil several criteria and allow us to either identify tracking images or to confidently release an image from the suspicion of being a tracking element. This is similar to a black- and whitelisting approach. Through the processing of the centrally provided CSS information for each image, an additional 24% of non-tracking images could be classified.

The *title* attribute can be used for various HTML elements. It was only used with one tracking image in our dataset and was in this case empty. Therefore, it can be assumed that if the title attribute with content is provided, the image is not a tracking image.

The attributes *vspace* and *hspace* define white spaces around images. The analysis showed that the tracking images have only zero as a value for both attributes. Therefore, all images that have a hspace or vspace value larger than zero could be classified as non-tracking images. Similarly, the attributes *align*, *id* and *usemap* only occur in non-tracking images. An image that uses any of the three attributes can be classified as a non-tracking image.

The *alt* and *class* attributes showed very mixed analysis results. We decided to leave them out since they could result in an overfitting optimisation of the detection model to the specific dataset, not leading to generally applicable results.

8.6.2 Reference Structure

The second category includes aspects that relate to the referencing link that points to the image, i.e. their URL. This category focuses on the textual and structural analysis of the reference. The usage of individualised links allows both the user and the e-mail to be identified. This is a necessary condition for tracking images.

Therefore, an important aspect is the detection of *individualised references*. In a first approach, we identified several aspects that distinguish tracking from non-tracking references. One example criterion involves combinations of letters, numbers, and again letters. Based on the insights from this initial analysis, we developed a scoring model indicating the likelihood that a link is individualised. Tracking references often fulfil the corresponding textual characteristics. Nonetheless, they are occasionally also encountered in non-tracking images. Therefore, several aspects need to be used in combination for detection.

The *wordlist* approach tries to identify tracking images by using a dictionary lookup to identify individualised parts of references. The idea is that non-tracking references (e.g., http://www.SLD.TLD/common/images/general/spacer.gif) usually contain a lot of common words that should be found in a dictionary, in contrast to tracking references that contain fewer or no common words. For the analysis, an English wordlist with 250,000 entries and a German wordlist with 190,000 entries were used. Each reference was split up into single parts that were checked in both wordlists. Afterwards, a measure was calculated that expresses the ratio of parts that were found in any of the dictionaries in relation to the total number of parts. However, it turned out that this approach did not provide results that clearly supported the decision process.

The *letter distribution* approach, similar to the wordlist, tries to identify how "normal" the image reference is, compared to typical distributions of letter occurrences within texts. Since links could be in both languages, distributions for the German (Beutelspacher, 2015) and English languages (Lewand, 2000) were considered. An analysis of the calculated measures shows that the deviations of the values for tracking images in relation to the rest of the images varied in both directions, higher as well as lower, while the deviation in the higher direction occurred more often. This deviation information can be used for tracking image identification.

Another aspect that could assist the identification of tracking images is the *similarity to other references*. Tracking references often distinguish themselves from other images in the mail with regard to their structure. A literature review on URL-similarity analysis revealed that various approaches have been developed. Often, the similarity of nodes (URLs) within a graph is used, which usually represents a subgraph of the World Wide Web (Benczúr et al., 2006; Cho et al., 1998; Lin et al., 2006; Maurer & Höfer, 2012; Menczer, 2004; Qi et al., 2007; Wu et al., 2012). Most approaches have been designed for hypertext web pages with two major requirements: the first is some machine interpretable text that can be used for textual word analysis, and the second involves hyperlinks that point to other web pages. However, both requirements are not fulfilled for image elements referenced in e-mails, which means that the graph-based approaches are currently not applicable for the given problem setting.

Therefore, another appropriate link-similarity measure was conceptualised. It is important to keep in mind that the measure should express structural similarity and not direct equality of each occurring character. In order to achieve this, the similarity of two links was defined as the amount of identical characters except digits, where the longer link is used as a comparison basis. If digits occur at the same position in each link, they are interpreted as equal characters,

regardless of whether or not they are actually identical.

The *keyword filter* evaluates whether the use of keywords is useful for the identification of tracking images. The idea is that specific words are only used in the context of tracking pixels, while some other keywords might only be used in non-tracking images. This approach is similar to the black-/whitelist approach. Since the goal was to identify keywords that are independent of specific senders and recipients, further filtering was necessary. Finally, a whitelist of 14 entries and a blacklist with 32 entries were created.

The *user-id as part of a reference* is an aspect that combines the advantages from the huge dataset and the reference analysis. The term 'user-id' is defined as a unique identifier which each tracking link of a specific sender contains and which is not part of any other non-tracking image link. It is assumed that the user identifier is included in every tracking link for the same recipient, while the e-mail or content identifier is different for each e-mail and would therefore only occur in a single mail. It turned out that user-ids could be determined for 41% of the tracking-mail senders.

The *text only* aspect analyses whether the image references are only composed of alphabetic letters, except special characters that are used for separation. Tracking links often contain randomly generated components including numbers for identification purposes. And, in fact, all tracking elements within our training dataset contained at least one digit.

The *file extension* aspect focuses on the file extensions of the images that are referenced within the dataset. While the majority of file extensions are used for both tracking and non-tracking images, some file extensions occur in only one category. For example, the extensions "cfm", "php" and "ssp" only occur in tracking images, while the extensions "io", "jpe", "ver" and "xiti" only occur in non-tracking images.

The *regular expression patterns* describe the structure of tracking-image references by means of regular expressions. Our analysis shows that tracking pixels from different senders are quite similar in terms of their structure. For the training dataset, 79 different types of patterns could be identified. After optimisation of the regular expressions, all tracking image references could be detected, while no non-tracking images matched. The regular expression approach would be sufficient to detect all tracking image references within the given dataset, but could result in heavy overfitting.

Further aspects that we analysed, but which did not improve results, included the amount of special characters, the actual number of numerical digits in links, and capital letters.

8.6.3 E-Mail Structure

The third category, e-mail structure, focuses on aspects that describe the occurrence of images based on their position within the structure of an e-mail. Furthermore, the number of occurrences is considered.

The *position* of images within e-mails is an important criterion. During the data analysis phase, it was noted that tracking pixels seem to occur often at the beginning or end of an e-mail. This seems reasonable, since they do not provide actual content. Another explanation might be the use of external services or software that just appends the tracking image to the top or end of an e-mail. Extended analysis reveals that, indeed, the vast majority of tracking images occurred at either the beginning or the end of an e-mail. If the first and last three images were always taken together, they accounted for 98.9% of all tracking images within the training set.

The second aspect is related to the number of occurrences of an image within an e-mail. The analysis of our data shows that no image that was referenced at least three times within the same mail was a tracking pixel. This information alone could be used to classify 39,994 (48%) image occurrences (3,134 different images) as non-tracking images.

8.6.4 Image Server

The fourth category, image server, describes aspects related to the server hosting the referenced images. The first aspect is the *occurrence of servers* within an e-mail. The relative occurrence of servers can provide valuable information with regard to tracking image detection. For example: If a company uses an external tracking provider, the regular images for the e-mail may be hosted by the company itself, while the tracking pixel is hosted by the tracking server provider on a different server. Therefore, the previously introduced term "main server" is useful. A short analysis showed that in the training dataset, 94% of the tracking images were not hosted on the main server. This supports the hypothesis that tracking images are usually not hosted on the main server if one exists.

The second aspect concerns the *location of the server*. First, the location of all image servers is determined via domain-name resolution and IP address geolocation (Wang et al., 2011). Then, the main server is used as a reference for all server assessments within a single mail. Location points are assigned as an approximation of the distance between the server in question and the main server. It turns out that more than two-thirds (68%) of the servers with the highest point score per email hosted tracking pixels.

8.6.5 Server Black-/Whitelisting

The fifth category is based on the entire dataset and distinguishes whether the image server in question is hosting only tracking images, only non-tracking images, or both. This idea of black- and whitelisting is borrowed from the detection and prevention of unsolicited emails (SPAM) (Cormack, 2006). For the application in our context, the elements of the lists are servers providing images that are referenced in the e-mails. This is an important difference from SPAM classification, where usually the sender or MTA is the object of investigation.

Our *blacklist* contains all servers that host tracking images but no non-tracking images. The second list is the *whitelist* with servers that only provide non-tracking images. The third case, *mixed hosts*, contains all servers that are part of neither the first nor the second list. This procedure was executed for the entire dataset. The majority of servers (57%) were part of

Table 8.5: Model summary and dataset dependency

Model Category	Information (Dataset Dependency)
Image Attributes	width, height, area, border, alt, style_blacklist, style_whitelist, vspace, hspace, title, class, align, id, usemap (all low dataset dependency)
Reference Structure	text-only (low), numbers_ratio (low), upper-letters_ratio (low), exceptional_reference (low), match_blacklist (medium), match_whitelist (medium), file_extension_blacklist (medium), file_extension_whitelist (medium), match_user-id (high), match_regular_expression (high)
E-Mail Structure	image_occurrence, internal_image, image_position (all low dataset dependency)
Image Server	server_location_points, main_server (both medium dataset dependency)
Server Black-/ Whitelisting	server_blacklist, server_whitelist (both medium dataset dependency)
Header Components	unsubscribe_link, reply_to, message_id, content_type, return_path, received_spf (all low dataset dependency)

the whitelist. A little more than one-third (34%) of the hosts were part of the blacklist. The remaining 9% of hosts were part of the mixed category, since they hosted both types of images. It has to be noted that the usage of servers could change over time. The approach is therefore only as up-to-date as the black-/whitelist itself, and regular updates should be ensured in order to minimise misclassifications.

8.6.6 Header Components

Any e-mail is composed of an e-mail body and an e-mail header. Usually, the e-mail header contains technical information and is not visible to the end user. The sixth category analyses fields of the e-mail header. Here, the header fields “list-unsubscribe”, “reply-To”, “message-ID”, “content-type”, “return-path”, and “received-SPF” have been selected based on initial tests. A method was developed which allowed already classified image links to be correlated with the six header fields. The majority of classifications could be realised through a customised text search in the list-unsubscribe header field. The header fields reply-to, content-type, return-path, and received-spf did not cause any misidentifications. Overall, 99.8% of the occurring matches were indeed tracking images. Therefore, header analysis proved very useful.

8.6.7 Detection Model Summary and Dataset Dependency

We now turn to discussing the dataset dependency of the detection model in order to estimate how suitable the individual attributes are for detecting tracking elements in new, unknown datasets. A distinction will be made through the association to one of the following groups: low, medium, or high dataset dependency. Table 8.5 gives an overview of the model aspects and their dataset dependency.

We assume that the aspects of *image attributes*, *e-mail structure*, *image server* and *header components* have low dataset dependency, since they represent properties that are characteristic for tracking images in general. The *black- / whitelisting* of image servers is assumed to be medium dataset dependent. Lists of tracking servers of external provider are applicable for all other e-mails that use the same provider.

The majority of *reference structure* attributes are assumed to have a low dataset dependency since they describe characteristics of tracking references that need to be fulfilled for the unique identification of e-mail and recipient. The reference keyword matching and the file extension analysis are assumed to have a medium dataset dependency, since they are relatively dependent on the data from which they have been derived, but are applicable to new datasets as well. The sender identifiers (user-id) are highly dependent on the dataset, since they were generated based on the dataset and are specific to the e-mail recipients. Even though the regular expressions are applicable to other e-mails as well, it is assumed that they are not characterising all possibly occurring tracking image references and might therefore mislead future detection processes for entirely different e-mails.

Taken together, the developed model seems to be very applicable to different e-mails. Since the aforementioned regular expression attribute seems to be highly dataset dependent and shows a high explanatory power at the same time, we do not consider the attribute for future classifications, since it might distort the classification results and accuracy evaluation for future e-mails. This step was taken in order to make the detection process more independent from the specific dataset that was used.

8.7 Validation

This section evaluates the classification accuracy and execution time of the proposed mechanism for detecting tracking images in e-mails. The approach consists of the detection model (described above) and a decision model, which aggregates the model aspects and forms an overall decision. In order to perform the validation, the total dataset was divided into a training set and a test set. With this separation, a distribution similar to the total dataset was for the goal, in particular with respect to the tracking elements. The separation of the 4,505 mails resulted in a distribution of 1,086 (24.1%) mails (26781 images) in the test set and 3,419 (75.9%) mails (83299 images) in the training set.

An artificial neural network (ANN) was used as a decision technique. Classification tasks are a popular application area of ANNs (Hastie et al., 2009). The attributes of the decision model are given as inputs to the ANN, which classifies whether or not the image in question is a tracking image. It is expected that the ANN approach benefits from its capacity to detect various complex patterns, as well as the ability to handle incomplete information (Hastie et al., 2009). For the given setting, multilayer perceptrons (MLPs) are appropriate. They can process categorical and numerical input and deliver categorical output values. They can also capture complex – nonlinear – relationships between inputs and outputs (Haykin, 1999). The network is generated using IBM Statistics in version 21. Since it was not our main goal to study the

Table 8.6: Confusion matrix for image classification

	Predicted Class	
	Tracking	Non-Tracking
Tracking	577 (TP)	0 (FN)
Non-Tracking	2 (FP)	26,202 (TN)

different possibilities for structural variations of the ANN, an automatic procedure was used for the network setup. As a training approach, the full batch method was applied in order to assure that the network was directly optimised in terms of a global optimum.

The *neural network* shows very good results regarding the classification of image elements in both training and test sets. The ANN was able to classify all images in the training set correctly. With regard to the classification of the unknown images in the test set, the overall accuracy was 99.99%. Table 8.6 shows the confusion matrix of the classification. A particularly important result is that no false negative classifications were generated that could threaten user privacy.

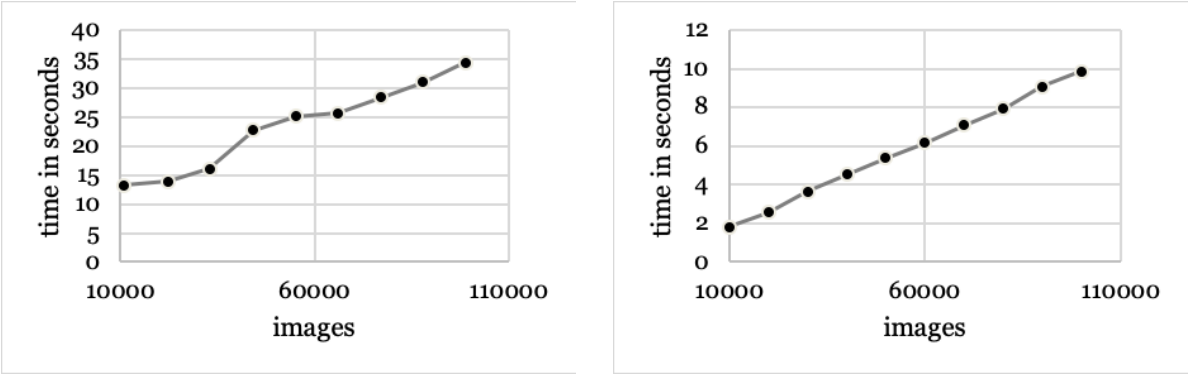
The execution time is also a major aspect with regard to the practicability of the prototypical solution. The process involves a setup and a usage phase. In the setup (training) phase, the model is built and optimised for the data basis. In the usage phase (classification), the model is applied to classify unknown images. The training procedure needs to be conducted only rarely and is independent from the classification. In our test environment, performance was determined separately for training and classification. The measurements were conducted under non-optimised conditions (virtual machine and limited main memory), leaving opportunities for performance improvements in subsequent applications. The measured times included the feature extraction, but without some fast preprocessing steps.

The training time (Figure 8.8a) for the ANN shows an approximately linear relationship to the number of images that were used for training. The *classification time* (Figure 8.8b) displays a nearly perfect linearity. Most importantly, classification was very fast with less than 0.2 ms per image. Full classification for a typical mail with 38 external images requires less than 7 ms. This result indicates a high practicality for real world application, even when many e-mails have to be processed.

8.8 Limitations

Finally, some important limitations of our research should be considered. The study focuses on *tracking images,* which represent a commonly applied and effective e-mail tracking mechanism. Other techniques such as tracking links have so far not been considered in the analysis and countermeasure design, even though similar approaches could be adopted.

The data used in this study only involve *professional e-mail newsletters,* which carry some important advantages for this research setting. Nonetheless, typical mail usage involves additional mail categories, especially individual mails. While we expect e-mail tracking to be less common, further studies are required on tracking mechanisms in these types of e-mails. Another aspect



(a) Training time

(b) Classification time

Figure 8.8: Performance for the ANN classifier

is the internationality of the dataset. Even though three representative countries were used for gathering the newsletters, tracking techniques adopted in other countries might slightly deviate and are currently not represented in the data set. The *location* analysis might not be highly accurate, since only free geolocation services were used for this study. Nonetheless, the country classification should be correct (Poese et al., 2011).

The *dataset dependency* of the tracking-image identification has already been discussed. Furthermore, this study uses data gathered in the year 2015. The results therefore reflect the current technological development at this time. It is very likely that the use of countermeasures will lead to an enhancement of tracking technologies, and will therefore make further advanced detection techniques necessary.

This study uses an *artificial neural network* for image classification. Alternative machine learning techniques could potentially show better results and should also be considered in further developments.

8.9 Conclusion

E-mail tracking can be used to gather sensitive information without user control, which raises several security and privacy concerns for end-users. Our empirical analysis of over 4,500 e-mails from the top 100 companies in Germany, Great Britain, and the United States showed that e-mail tracking is widely applied in all of these countries. Out of all the e-mails that could potentially contain tracking elements, 71% actually used tracking images. The tracking quota of German mails was the highest at 88%, followed by the mails from the US with 63% and UK with 62%. While the tracking in German mails heavily relied on internal tracking, the tracking mails from the US mostly relied on external providers. Our evaluation of current countermeasures showed that currently no general, reliable, and sufficient protection against e-mail tracking exists. This constitutes a demand for a universal and reliable protection method.

As a first step in the direction of countermeasure realisation, several concepts have been developed and discussed in this study. With regard to end-user demands, the *identify & block* solution seems to be the most suitable, since it aims at selectively identifying and blocking

tracking images while permitting other referenced images. Since e-mail tracking images so far do not provide the receiver with any content and are typically invisible, this solution does not reduce functionality or usability for the recipient.

Based on the analysis of our large empirical dataset with of over 110,000 images, a detection model was developed which encompasses six categories of important aspects that are useful for classifying unknown images. An artificial neural network was created based on the detection model and used as the decision technique for image classification.

Finally, the usefulness of our approach was evaluated using experiments on a test dataset. The experimental results showed that the neural network classified 99.99% of the images correctly and that no problematic false negatives occurred. Moreover, the execution speed of our classification algorithm was fast, indicating its practical usefulness for future work on a fully implemented countermeasure solution.

Bibliography

Agosti, M., & Di Nunzio, G. M. (2007). Gathering and Mining Information from Web Log Files, In *Digital Libraries: Research and Development*, Berlin, Heidelberg, Springer. https://doi.org/10.1007/978-3-540-77088-6_10

Alsaid, A., & Martin, D. (2003). Detecting Web Bugs with Bugnosis: Privacy Advocacy through Education (R. Dingledine & P. Syverson, Eds.). In R. Dingledine & P. Syverson (Eds.), *Privacy Enhancing Technologies*, Berlin, Heidelberg, Springer. https://doi.org/10.1007/3-540-36467-6_2

Benczúr, A. A., Csalogány, K., & Sarlós, T. (2006). Link-based Similarity Search to Fight Web Spam, In *Adversarial Information Retrieval on the Web AIRWeb*.

Beutelspacher, A. (2015). *Kryptologie*. Wiesbaden, Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-05976-7

Bonfrer, A., & Drèze, X. (2009). Real-time evaluation of e-mail campaign performance. *Marketing Science*, *28*(2), 251–263. https://doi.org/10.1287/mksc.1080.0393

Bouguettaya, A., & Eltoweissy, M. (2003). Privacy on the web: Facts, challenges, and solutions. *IEEE Security & Privacy Magazine*, *1*(6), 40–49. https://doi.org/10.1109/MSECP.2003.1253567

Burkell, J., & Fortier, A. (2013). Privacy policy disclosures of behavioural tracking on consumer health websites. *Proceedings of the American Society for Information Science and Technology*, *50*(1), 1–9.

Campaign Monitor. (2010). How Do I Create a Printer-Friendly Email Newsletter?

Cho, J., Garcia-Molina, H., & Page, L. (1998). Efficient crawling through URL ordering. *Computer Networks and ISDN Systems*, *30*(1), 161–172.

Cormack, G. V. (2006). Email spam filtering: A systematic review. *Foundations and Trends® in Information Retrieval*, *1*(4), 335–455.

Dobias, J. (2011). Privacy Effects of Web Bugs Amplified by Web 2.0, In *Privacy and Identity Management for Life*, Berlin, Heidelberg, Springer. https://doi.org/10.1007/978-3-642-20769-3_20

Evans, M., & Furnell, S. (2003). A model for monitoring and migrating web resources. *Campus-Wide Information Systems*, *20*(2), 67–74. https://doi.org/10.1108/10650740310467763

Evers, J. (2006). How HP Bugged E-Mail: Commercial Online Service Was Used to Track E-mail Sent To a Reporter in Hewlett-Packard's Leak Probe, Investigator Testifies. *CNET*.

Fabian, B., Bender, B., & Weimann, L. (2015). E-Mail Tracking in Online Marketing: Methods, Detection, and Usage, In *12th International Conference on Wirtschaftsinformatik*, Osnabrück, Germany.

Fonseca, F., Pinto, R., & Meira, W. (2005). Increasing User's Privacy Control through Flexible Web Bug Detection, In *Third Latin American Web Congress*, Buenos Aires, Argentina, IEEE. https://doi.org/10.1109/LAWEB.2005.19

Fourie, I., & Bothma, T. (2007). Information seeking: An overview of web tracking and the criteria for tracking software. *ASLIB Proceedings*, *59*(3), 264–284. https://doi.org/10.1108/00012530710752052

Harding, W. T., Reed, A. J., & Gray, R. L. (2001). Cookies and web bugs: What they are and how they work together. *Information Systems Management*, *18*(3), 17–24.

Hasouneh, A. B. I., & Alqeed, M. A. (2010). Measuring the effectiveness of e-mail direct marketing in building customer relationship. *International Journal of Marketing Studies*, *2*(1), 48–64. https://doi.org/10.5539/ijms.v2n1p48

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The Elements of Statistical Learning* (Second). New York, Springer.

Haykin, S. S. (1999). *Neural Networks: A Comprehensive Foundation* (Second). Upper Saddle River, Prentice Hall.

Javed, A. (2013). POSTER: A Footprint of Third-Party Tracking on Mobile Web, In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, Berlin, Germany, ACM. https://doi.org/10.1145/2508859.2512521

Jensen, C., Sarkar, C., Jensen, C., & Potts, C. (2007). Tracking Website Data-collection and Privacy Practices with the iWatch Web Crawler, In *Proceedings of the 3rd Symposium on Usable Privacy and Security*, Pittsburgh, PA, USA, ACM Press. https://doi.org/10.1145/1280680.1280686

Krishnamurthy, B., & Wills, C. (2009). Privacy Diffusion on the Web: A Longitudinal Perspective, In *Proceedings of the 18th International Conference on World Wide Web*, Madrid, Spain, ACM Press. https://doi.org/10.1145/1526709.1526782

Leon, P., Ur, B., Shay, R., Wang, Y., Balebako, R., & Cranor, L. (2012). Why Johnny Can't Opt Out: A Usability Evaluation of Tools to Limit Online Behavioral Advertising, In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Austin, Texas, USA, ACM Press. https://doi.org/10.1145/2207676.2207759

Lewand, R. (2000). *Cryptological Mathematics*. Washington, DC, Mathematical Association of America.

Lin, Z., Lyu, M. R., & King, I. (2006). PageSim: A Novel Link-based Measure of Web Page Similarity, In *Proceedings of the 15th International Conference on World Wide Web*, ACM.

Litmus Labs. (2015). Email Client Market Share: Email Client Usage Worldwide.

Martin, D., Wu, H., & Alsaid, A. (2003). Hidden surveillance by web sites: Web bugs in contemporary use. *Communications of the ACM*, *46*(12), 258. https://doi.org/10.1145/953460.953509

Maurer, M.-E., & Höfer, L. (2012). Sophisticated Phishers Make More Spelling Mistakes: Using URL Similarity Against Phishing, In *Cyberspace Safety and Security*, Springer.

Menczer, F. (2004). Combining Link and Content Analysis to Estimate Semantic Similarity, In *Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters*, ACM. https://doi.org/10.1145/1013367.1013521

Moscato, D. R., Altschuller, S., & Moscato, E. D. (2013). Privacy policies on global banks' websites: Does culture matter? *Communications of the IIMA*, *13*(4), 91–109.

Musciano, C., & Kennedy, B. (2006). *HTML & XHTML: The Definitive Guide*. O'Reilly Media, Inc.

Nakatani, K., & Chuang, T.-T. (2011). A web analytics tool selection method: An analytical hierarchy process approach. *Internet Research*, *21*(2), 171–186. https://doi.org/10.1108/10662241111123757

Peffers, K. E. N., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77.

Poese, I., Uhlig, S., Kaafar, M. A., Donnet, B., & Gueye, B. (2011). IP geolocation databases: Unreliable? *ACM SIGCOMM Computer Communication Review*, *41*(2), 53. https://doi.org/10.1145/1971162.1971171

Premkumar, G., & Roberts, M. (1999). Adoption of new information technologies in rural small businesses. *Omega*, *27*(4), 467–484.

Qi, X., Nie, L., & Davison, B. D. (2007). Measuring Similarity to Detect Qualified Links, In *Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web*, ACM.

Sipior, J. C., Ward, B. T., & Mendoza, R. A. (2011). Online privacy concerns associated with cookies, flash cookies, and web beacons. *Journal of Internet Commerce*, *10*(1), 1–16. https://doi.org/10.1080/15332861.2011.558454

Waisberg, D., & Kaushik, A. (2009). Web analytics 2.0: Empowering customer centricity. *The Original Search Engine Marketing Journal*, *2*(1), 5–11.

Wang, Y., Burgener, D., Flores, M., Kuzmanovic, A., & Huang, C. (2011). Towards Street-Level Client-Independent IP Geolocation, In *Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation NSDI'11*, USENIX Association Berkeley, CA, USA. https://doi.org/10.5555/1972457.1972494

Wu, Y.-J., Huang, H., Hao, Z.-F., & Chen, F. (2012). Local community detection using link similarity. *Journal of Computer Science and Technology*, *27*(6), 1261–1268.

Chapter 9

Track and Treat: Usage of E-Mail Tracking for Newsletter Individualization

Publication
Bender, B., Fabian, B., Haupt, J., Lessmann, S., Neumann, T., & Thim, C. (2018). Track and Treat — Usage of E-Mail Tracking for Newsletter Individualization. Proceedings of the 26th European Conference on Information Systems (ECIS'18).

Abstract
E-Mail tracking mechanisms gather information on individual recipients' reading behavior. Previous studies show that e-mail newsletters commonly include tracking elements. However, prior work does not examine the degree to which e-mail senders actually employ gathered user information. The paper closes this research gap by means of an experimental study to clarify the use of tracking-based information. To that end, twelve mail accounts are created, each of which subscribes to a pre-defined set of newsletters from companies based in Germany, the UK, and the USA. Systematically varying e-mail reading patterns across accounts, each account simulates a different type of user with individual reading behavior. Assuming senders to track e-mail reading habits, we expect changes in mailer behavior. The analysis confirms the prominence of tracking in that over 92% of the newsletter e-mails contain tracking images. For 13 out of 44 senders an adjustment of communication policy in response to user reading behavior is observed. Observed effects include sending newsletters at different times, adapting advertised products to match the users' IT environment, increased or decreased mailing frequency, and mobile-specific adjustments. Regarding legal issues, not all companies that adapt the mail-sending behavior state the usage of such mechanisms in their privacy policy.

9.1 Introduction

E-mail tracking encompasses methods for gathering information regarding an individual user's reading behavior. Previous studies show that professional e-mail senders routinely embed tracking elements in newsletters and other marketing communication (Fabian et al., 2015). Since tracking is often conducted without consent of the tracked individual, such practices raise ethical and privacy concerns, especially because the majority of users is unaware of the possibility to track e-mail reading behavior (Thode et al., 2015).

E-mail tracking approaches split into tracking links and tracking images. The former use em-

bedded references to collect information once a user opens the link in an e-mail. In this sense, the tracking link approaches requires active participation from the user in the form of clicking a link. Tracking images are images embedded in HTML-based e-mails, which e-mail clients fetch from a (tracking) server once a user opens an e-mail. They facilitate data collection regarding the user reading behavior *without the recipient's permission* (Bender et al., 2016), thus exacerbating their threat to data privacy and justification from an ethical point of view.

Previous research focuses on the prevalence of e-mail tracking (Fabian et al., 2015) and the detection of potential tracking images within e-mail communication (Bender et al., 2016). A limitation of prior work lies in its focus on the detection of elements that *potentially* facilitate tracking. For example, embedding a tracking image in an e-mail fulfills the technological prerequisites to track whether a user opens an e-mail. However, confirming the presence of tracking elements in e-mails does not clarify the extent to which senders actually process and employ the information they can potentially gather. Examining the actual use of tracked information is the goal of this paper. In particular, this study clarifies whether e-mail senders adjust their communication policies in response to user data gathered through tracking. In line with prior work, we focus on professional e-mail newsletters because such communication serves a marketing goal and thus incentivizes senders to individualize e-mail messages.

To examine the use of tracking data by commercial mail senders, we design an experiment with twelve e-mail accounts, each of which simulates a specific type of user with individual e-mail reading behavior. We ensure that behavioral differences across user accounts are easy to track by means of tracking images (Suneetha & Krishnamoorthi, 2009). Each account subscribes to the same set of professional newsletters, which we gather from companies of various industries. We concentrate on German, British and US companies to evaluate cross-country differences related to different regulations and legal restrictions.

To the best of our knowledge, this is the first experimental study that provides evidence that companies actually use the data they collect through e-mail tracking to adjust marketing communication on an individual level. Our analysis also reveals that a fraction of companies employ personal response data to individualize the frequency, timing and content of marketing communication. These results confirm that data collection, storage and analysis on the personal level takes place and emphasizes the need for additional research regarding the extent of identified privacy risks and the development of efficient protection strategies.

We organize the paper as follows. The next section discusses prior work on e-mail tracking and related tracking technologies. Section 3 elaborates on our experimental design. We then present and discuss results in Section 4. Section 5 concludes the paper.

9.2 E-Mail Tracking Fundamentals

This section discusses the process of e-mail tracking and its technological fundamentals in the form of tracking links and images. *Tracking links* are hyperlinks in an e-mail that are augmented with identifiers, which are not part of the reference but convey information about interaction

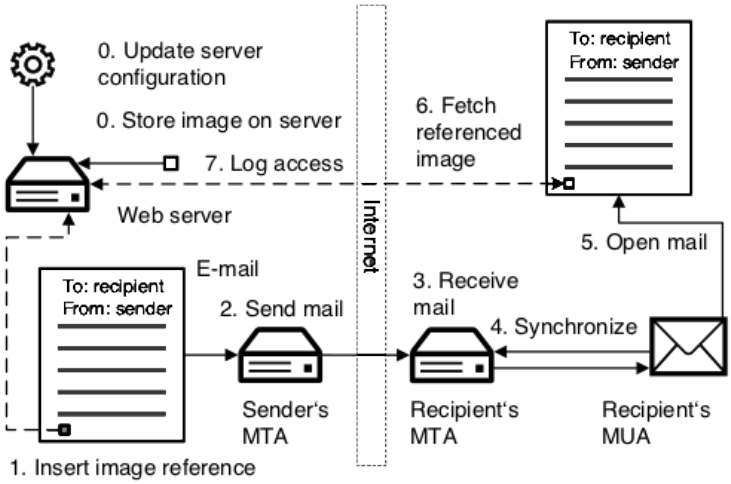


Figure 9.1: E-mail tracking process (see Bender et al. (2016))

with the link. In particular, tracking links can include a unique identifier that allows to detect and log whether an individual e-mail recipient follows the link (Fabian et al., 2015). Technically, this is typically realized using an individual link for every recipient to be able to detect any website request using web server analytics (Agosti & Di Nunzio, 2007) or a redirection service (Nikiforakis et al., 2014). The latter also facilitates matching the browsing behavior on the target page with an e-mail recipient through the identifier transmitted via the referrer URL in the specialized link (Jin et al., 2010).

Tracking images are external image references within HTML e-mails that contain identifying information. Figure 9.1 depicts the tracking process for e-mails that reference external image resources. The sender prepares an HTML e-mail including an image reference augmented by information on the identity of the receiver and the content of the e-mail. After the e-mail is sent, it passes several mail transfer agents (MTAs) until it reaches the receiver's MTA. Next, the recipient opens a mail client, which synchronizes the local mail repository with the newest version of the recipient's MTA. When the recipient opens the e-mail with a tracking image, the mail client requests the image from the referenced destination. The web server logs this request and provides the image to the recipient's mail client. Finally, analysis of the server log files provides detailed insights on the recipient's e-mail reading behavior.

9.2.1 Related Literature

E-mail tracking can be interpreted as the application of common web tracking mechanisms in the e-mail context. In the following, we discuss 1) relevant web tracking studies and 2) studies specifically related to e-mail tracking.

The tracking of web users has been an activate topic of research, see the recent surveys by Bujlow et al. (2017) and Ermakova et al. (2018) as well as individual studies (Acar et al., 2014; Bouguettaya & Eltoweissy, 2003; Englehardt & Narayanan, 2016; Ermakova et al., 2017; Evans & Furnell, 2003; Gomer et al., 2013; Hamed et al., 2013; Han et al., 2012; Libert, 2015; Roesner

et al., 2012; Schelter & Kunegis, 2016a, 2016b). The web-browsing behavior of online users is considered a worthwhile source for detailed profiling (Falahrastegar et al., 2016; Mitchell, 2012) to improve commercial activities such as targeted advertising (O'Connell, 2014; Roesner et al., 2012). Enabled by a variety of techniques (Bujlow et al., 2017; Sanchez-Rola et al., 2017), web tracking has become ubiquitous (Ermakova et al., 2017; Roesner et al., 2012; Schelter & Kunegis, 2016a, 2016b) on single sites, but also across websites and even across devices (Brookman et al., 2017; Falahrastegar et al., 2016; Gomer et al., 2013; Mayer & Mitchell, 2012). Some articles have analyzed the methods and extent by which relevant information can be extracted from tracking data (Bujlow et al., 2017; Suneetha & Krishnamoorthi, 2009). Besides targeted advertising (Parra-Arnau, 2017; Sanchez-Rola et al., 2017), web tracking can be applied for personalization, advanced web-site analytics, and social network integration (Mayer & Mitchell, 2012; Roesner et al., 2012; Sanchez-Rola et al., 2017).

For online users, web tracking practices can result in increased online privacy risks (Jin et al., 2010; Mayer & Mitchell, 2012; D. R. Moscato & Moscato, 2009; Roesner et al., 2012), including price discrimination, government surveillance, and identity theft (Bujlow et al., 2017). The extent to which tracking is made transparent within the privacy policies of business to consumer companies depends on user expectations (D. R. Moscato et al., 2013).

E-mail tracking, i.e. the use of web tracking methods in e-mail communication, has become a growing concern in scientific literature as well as in the public press. A description of techniques for extracting user information from e-mails is given by Foulger et al. (2008) and Cselle et al. (2007). As discussed by Fabian et al. (2015), e-mail tracking allows the collection of detailed information on individual reading behavior without explicit consent of the user. In this regard, tracking images represent a more severe privacy issue since information is collected automatically when an e-mail is opened, whereas tracking links require active clicking on the referenced content. Bender et al. (2016) provide a first international study regarding the use of e-mail tracking in commercial newsletters and focus on the conceptualization of potential countermeasures.

There are studies that highlight some functional advantages of e-mail tracking, e.g., that the basic structure of the e-mail service does not allow a sender to be certain that a message is really delivered to the right receiver (Oppliger, 2007). Schmidt (2013) discusses the usage of tracking images and evaluates current protection through commonly used e-mail software for personal use. The information that can be collected comprises primary information that can be gathered directly from the tracking server logs, and secondary information, based on additional resources to enhance and combine with the primary information. Examples of primary information include the time or the client's user-agent string that was used to request the image. Examples for secondary information are the location from which the e-mail is retrieved as well as potentially a user's affiliation, or if an e-mail has been printed or forwarded (Bender et al., 2016). The combination of information allows building a profile of the individual user's behavior.

An important aspect that distinguishes e-mail tracking from general web tracking techniques

is that the collected data is not anonymous, since it can be directly attributed to an e-mail recipient identified by a unique e-mail address (Jin et al., 2010). Since email addresses often contain the name of the individual and the name of an affiliated institution in the domain and are often used to sign in on several websites, some of which may require personal information, they facilitate the identification of individuals to a larger extent than web tracking.

9.3 Study Design

We conduct a controlled experiment by simulating user interaction with marketing newsletters in order to evaluate whether e-mail senders vary their communication and sending policies depending on the recipient's reading behavior. Using a set of artificial user accounts allows us to minimize confounding factors by standardizing user characteristics. This section describes the experimental setup used for data collection and the user behavior profiles.

To collect data, we set up twelve up e-mail accounts on Gmail. Ten accounts simulate a specific, consistent user behavior. The remaining two accounts do not conduct any activity to allow comparison and validation of the results. We create all user identities to be older than 21 years to eliminate potential restrictions in the offerings and choose user birthdays to be outside the data-gathering period to eliminate bias from potential birthday related offerings. Given our focus on tracking, all identities share the same gender (male) to avoid gender-specific offerings in view of the content comparison. Other personal information required during newsletter registration is held constant over identities and matched to characteristics we expect for subscribers of each company. For example, we use country-specific addresses to prevent a potential relocation to another subsidiary of the company.

For each newsletter subscription, we ensure a 'clean' browser environment to prevent potential linking between the accounts and the deduction of preferences from the browser history, which is a common practice in web tracking (Nikiforakis et al., 2014). For example, we delete cookies, history and form entries, etc. from the web browser cache and begin each registration process from a new browser session. In addition, we ensure location-specific IP addresses within the subscription process. Finally, we limit the information provided to companies during newsletter signup to mandatory entries. This helps preventing content variation based on preferences or attributes given during the subscription. In case where such information was mandatory, we provided the same data for all accounts.

We register each mail account for the same set of commercial newsletters selected from the largest e-commerce companies based in Germany, the United Kingdom, and the United States of America. Large companies are likely to have knowledge as well as the resources to employ individual targeting and complex analytic solutions. This study focuses on companies from one industry, online retail, for several reasons. First, online retail specializes in digital business. Therefore, we expect companies to be well developed with regard to technological possibilities in general and technologies to enhance customer-centric processes in particular. Second, a large product portfolio simplifies segmentation and individualization of offerings compared to other industries (e.g., public transport, manufacturers, etc.). Third, personalization and customer

targeting are established success factors in online retail (Golrezaei et al., 2014). It is thus plausible to expect e-tailors to be pioneers in personalization.

Within retail, we consider six areas: clothing, electronics, general retails, home goods, supermarket and tourism. Within these areas, we select on successful and large business-to-consumer retailers as determined by means of country-specific rankings based on revenue or sales (Germany: EHI Retail Institute (2019), USA: eMarketer, cited in Zaczkiewicz (2016), UK: UK (2016)). The rationale for this selection is that large retailers are more likely to have the resources and know-how to engage in tracking, targeting, and personalization. The listed companies were assigned to each of the six retail categories, if applicable. Trading companies without a retail focus were excluded from the study in order to ensure a defined and comparable sample of companies with an incentive for newsletter personalization. Furthermore, globally active trading companies, e.g. Amazon, were excluded from the study, since the attribution to a single region is imprecise and might distort the country-specific results.

In total, each of the 12 e-mail accounts subscribes to 52 company newsletters. E-mails are collected for a ten-week period from the 12th to the 21st calendar week of 2017. We argue this a reasonable time span for newsletter senders to collect user information and to adjust or individualize e-mails.

Simulating different user behaviors and retrieving e-mails to access referenced images requires a dedicated and customizable software. We have developed a corresponding system using the Java programming language. Java is a suitable choice because it features various easy-to-use components such as JavaMail for mail access and JSoup for parsing XML-based files like HTML-based mails that jointly provide the required functionality. All information gathered for the experiment is stored in a relational database. Importantly, to simulate different scenarios, all images within one mail are fetched according to the individual user/account profile.

Table 9.1: Simulated behavior of e-mail accounts in the experiment

	Factor			
Account	Frequency	Reading time	Device type	Location
1	1/day	random	Windows	Germany
2	3/day	random	Windows	Germany
3	1/day	fixed time (1pm)	Windows	Germany
4	1/day	fixed time (10pm)	Windows	Germany
5	1/day	3 minutes after reception	Windows	Germany
6	random	random	Windows	Germany
7	1/day	random	Windows	USA
8	1/day	random	OSX	Germany
9	1/day	random	Android	Germany
10	1/day	random	iOS	Germany
11	never	none	-	-
12	never	none	-	-

Since the study aims to evaluate the use of information gathered through tracking images during

e-mail reading, the experiments need to simulate relevant user behavior. For the e-mail tracking process (Figure 9.1) it is essential to fetch external referenced content during the reading process. Thereby, the tracker can use all the information available during the image request to build a profile. From a conceptual point, we divide the information in infrastructural and behavioral aspects that might influence the newsletter targeting. Infrastructural information such as the devices used are typically static thus making it easier to conduct corresponding targeting activities. The behavioral aspects are more dynamic and it is therefore more complex to deduce corresponding targeting activities.

To implement the behavioral user profiles, we develop a separate request function for every test account. Each account can have its own settings for the IP address, user agent string and predefined execution time. To allow for simultaneous image requests of different test accounts, a multithreading procedure has been employed to comply with concurrency requirements. We use predefined timers to start the respective threads, which helps steering the exact time sequences for image requests.

Table 9.1 gives the experimental factors for each account. To simulate different user behavior, we vary the time and frequency of e-mail access and the device type and location.

Accounts open e-mails with a reading frequency fixed at once a day with the exception of account #2, which opens e-mails three times a day, and account #6, which opens each e-mail at a random time and frequency, but at least once. The time at which e-mails are opened is randomly drawn from a uniform distribution over the minutes of the day per account and day for most accounts. Accounts #3 and #4 open all new e-mails at a fixed time of the day at 1 a.m. and 10 p.m., respectively. Account #5 opens e-mails three minutes after they are received.

We fix the device type for each account through manipulation of the user-agent string to the desktop (Windows/OSX) or mobile (Android/iOS) operating system of the most common vendors Microsoft and Apple, respectively. Each account accesses e-mails and external content through one of two proxy servers to fix the location derivable from the IP address. We use the same proxy server with a static IP address for the duration of the experiment. Locations are either a German university or a school located in Hanford, California, for account #7.

9.4 Study Results

We begin the presentation and discussion of empirical results with reporting descriptive statistics related to the newsletter e-mails gathered through the user accounts. During the collection period, we receive 12,404 valid e-mails in total, of which 12,346 are in HTML. The HTML e-mails can include tracking images. Not all companies started delivering newsletters. In total, 44 out of the 52 companies sent newsletters and the sending behavior differs across newsletters. Most of the e-mails come from German companies (45.6 %) whereas the share of newsletters from the USA and UK are 34.4 % and 20 %, respectively. Newsletter shares across industries per country show that in Germany the dominating industry is supermarkets, in contrast to both other countries, where the general retail and home goods companies use e-mail marketing

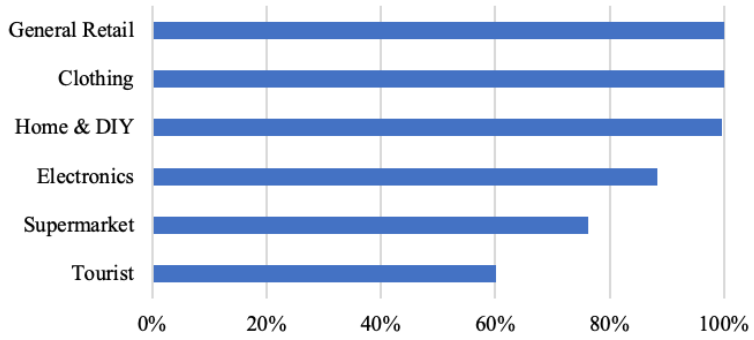


Figure 9.2: Tracking rate for different trading industries

much more often. The clothing sector is similarly prominent in all three countries.

We employ the detection model of Bender et al. (2016) to identify tracking images. In all countries, the prevalence of tracking is comparable and at a high level. Newsletters from German companies have the lowest amount with 85.68 % of all e-mails containing at least one tracking image. 93.76% e-mails from the UK contain tracking images and 99.48% from the US. Examining the share of tracking e-mails across industries reveals that all newsletters from the General Retail and Clothing and 99% from the Home Goods sector contain tracking images (see Figure 9.2). To a lesser degree, 88% of electronics newsletters and 76% of supermarket e-mails contained tracking images, while touristic newsletters showed the lowest tracking rate with 60%.

In the following subsections, we evaluate each experimental factor varied in the experimental design. We begin with an analysis of the overall number of mails received per account. Afterwards, location specific adjustments based on the offsite account are evaluated. We then evaluate content variation between the newsletters in the different accounts. Finally, results regarding varying sending behavior for the individual simulated accounts are discussed.

9.4.1 Amount of E-Mails Received

Analyzing the number of received e-mails for each account provides a first indication of differential sender behavior. The most remarkable aspect is that both validation accounts received substantially less e-mails than all other test accounts. This is a clear indication that companies observe the opening rates of subscribers and adjust the sending behavior accordingly.

Within the other accounts, the number of received e-mails ranges from 1,026 to 1,094 messages, with an average of 1,063 mails per account. The noticeably smaller number for test account #7 is due to one newsletter that, for unknown reasons, has not been delivered to test account #7 while being active in all other accounts. To use a consistent and comparable data basis, we exclude this newsletter, i.e. 21 e-mails per account, from the subsequent analysis. The other differences in the number of received e-mails between accounts can be attributed to small divergences in the number of mails sent across all companies (e.g., as opposed to a large deviation in the sending behavior of a small number of companies). During the data-gathering period,

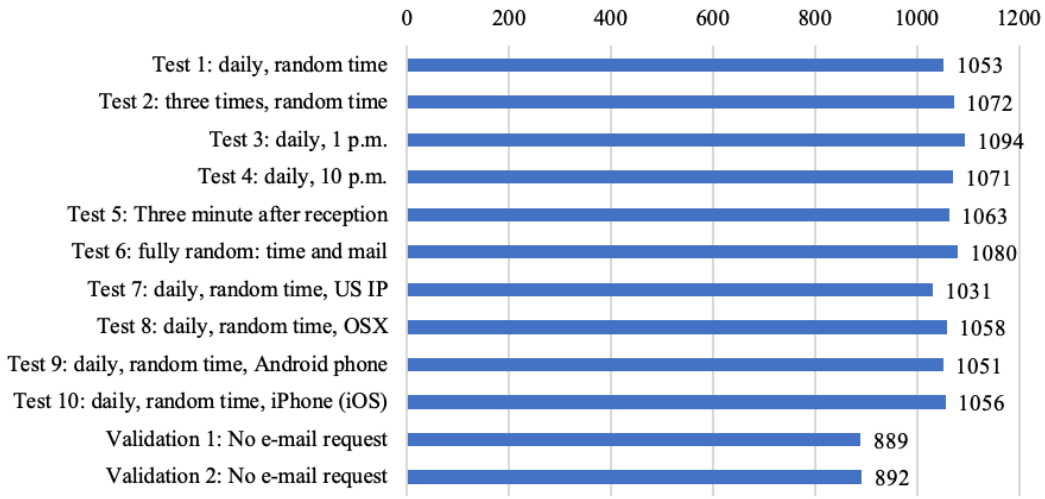


Figure 9.3: Received e-mails per account

we observe a slight increase in the number of received mails for all test accounts, but not for the validation accounts.

We observe a significant difference in the number of received e-mails between the ten treatment and two control accounts (Welsh t-test, df = 9.9998, $p < 0.001$). The difference provides evidence that companies use e-mail tracking information to adjust their communication policy. Further analysis of sending patterns reveals that companies stopped sending e-mails completely after no e-mail openings were tracked, with two companies stopping after only one unopened e-mail and two companies stopping after two to three e-mails. One company explicitly acknowledged the observed behavior after sending four unopened newsletters to the validation accounts by sending an e-mail with the message "we miss you" and special promotions. None of the test accounts receives a comparable message. We take the retention offer as strong evidence that the confirmation of e-mail openings provided by tracking images is used to target customer individually.

Interestingly, we also observe a novel newsletter in the data. In particular, one company not within the experiment selection started sending messages to the test accounts but not the validation accounts, without explicit newsletter subscription by any account. An analysis of the company's affiliation revealed that the company is affiliated with a company that the accounts subscribed to. We interpret these findings to confirm that i) e-mail addresses are transferred to the subsidiary and ii) that the addresses are further qualified with information collected through e-mail tracking. We interpret the selective behavior to show that the company uses observed reading behavior to select only active accounts for transfer. Further research is necessary to establish if the data transferred to the subsidiary includes information on the individual reading behavior in addition to the e-mail address.

We observe weakly significant variation in the amount of e-mails received between mobile and desktop users (Welch t-test, df = 7.67, p-value = 0.038), with the mobile accounts receiving less e-mails.

9.4.2 Location-Specific Adjustments

We simulate one user to open newsletters from a different country to test for location-based targeting. Since global companies have local subsidiaries that could target customers directly, we expect to observe adjustment of the sender or localized communication content. However, the data do not indicate major differences in the sending behavior. None of the issuing companies changes its top-level domain or the address from which newsletters are sent in response to test account #7 opening each e-mail from a U.S. IP address. Possible reasons for companies to ignore the IP location are that IP addresses can convey false information, e.g. if a mail proxy server or VPN is in use, and that location information may be temporary, e.g. when a user opens mail on holiday.

9.4.3 E-Mail Content Adjustment

Beyond the adjustment of the sending schedule, e-mails can be personalized by changes in e-mail phrasing, formatting and content. We therefore go on to compare the corresponding e-mails across the different accounts for their body length, textual content and image URLs (excluding the tracking images).

Figure 9.4: Example mail from electronic retailer. Test account #10 (right) receives information on considerably more Apple products than the other accounts (left)

The length of the text in e-mails received by each account could provide a first indication that systematic adjustment of e-mail content takes place. We conduct an analysis-of-variance (ANOVA) on e-mail length in characters but are unable to reject the null hypothesis of e-mail length with equal mean for all accounts ($F(11, 12132) = 0.84$, $p = 0.60$). Nevertheless, we often find substantial variation between accounts for a single e-mail. After inspection of these differences, we propose that A/B testing to be the main cause for the observed variance, where senders are using different design versions of the same e-mail to test the effectiveness of design choices. In some cases, several accounts received the exact same version of one e-mail while the other accounts received a different version.

In addition to the text of e-mails, marketing practice suggests to prefer short subject lines when targeting mobile customers due to their limited screen size. The difference in the average length of the e-mail subject for mobile compared to desktop devices is small at 0.8 characters and statistically insignificant (Welsh t-test, df = 9.05, p-value = 0.14).

Similarly, we frequently observe the use of different image versions or different icons but are unable to determine any structure within the deviations. Other differences in the e-mails are variations of personalized promotion codes or the recipients' e-mail address mentioned in the fine print.

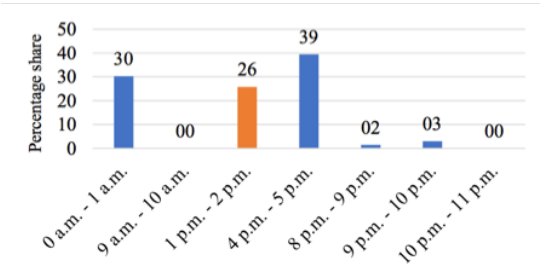
However, one electronic retail company specialized on computer, notebooks, mobile devices and peripherals adjusts marketing content based on the user's device. We observe that test account #10, which simulates an iPhone receives mails with substantially more Apple-related products than the other accounts over the full observation period. An example of the typical product offering for a comparison account (left) and account #10, which simulates the iPhone client, is presented in Figure 9.4. To test these observations, we identify the keywords Apple, iPhone, and MacBook, which occur substantially more often for this account than for the comparison accounts. While the iPhone account receives 576 mails containing Apple keywords, the account simulating the Apple laptop (#8) receives 517, which is slightly above the average of comparison accounts at 509.1. The difference in the average count of keywords between the accounts simulating an Apple system (#8 and #10) and all other accounts is not significant (Welsh t-test, df = 1.7729, p-value = 0.40).

We conclude that, even though variation could be observed, we are unable to identify statistically significant patterns or systematic variation and find no evidence that companies personalize the content of the newsletter based on information collected through e-mail tracking.

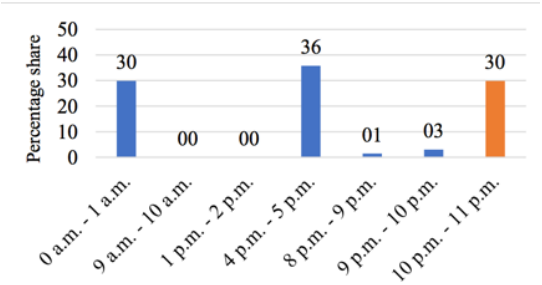
9.4.4 Sending-Time Adjustments

E-Mail users read their mails at different times. Some may read mails occasionally within usual business hours only, whereas others check mails more frequently. Data on a user's reading behavior may convey information regarding her digital media usage and daily routine, which are valuable insights for marketing. To examine whether e-mail senders adjust their communication according to the reading times of recipients, test accounts #3-5 simulate different reading styles. Account #3 read mails at 1 p.m., while test account #4 reads at 10 p.m. Test account #5 simulates frequent e-mail checking upon notification that an e-mail has been received.

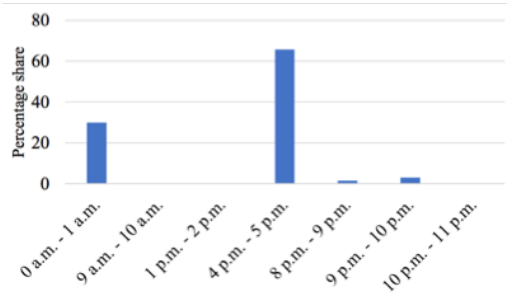
Taking the whole set of mails into account, we observe that only a single, very active company adjusts sending times in response to recipient behavior (Figure 9.5). Figure 9.5c and 9.5d show the e-mails from this company received by the validation accounts without e-mail access behavior. The mails that these receive as well as their time distribution are identical. The time distribution of email received by test account #3 (Figure 9.5a) and test account #4 (Figure 9.5b) differs substantially from the validation accounts. Although both accounts receive the same amount of 100 e-mails, the time at which these are sent differs and matches the different



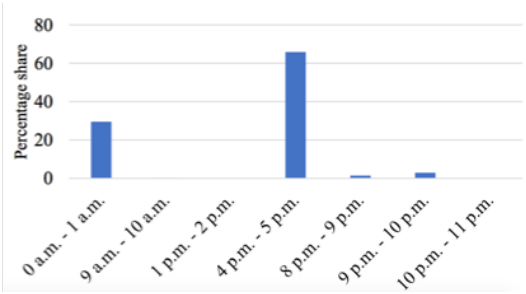
(a) Test account #3 (reading at 1 p.m.)



(b) Test account #4 (reading at 10 p.m.)



(c) Validation account #1 (e-mails not opened)



(d) Validation account #2 (e-mails not opened)

Figure 9.5: Number of e-mails received by a time-adjusting company per hour of day (periods with no deviation omitted for clarity)

reading (time) preferences simulated by the accounts. While both validation accounts and test account #4 receive no mails between 1 and 2 p.m., test account #3 that read mails at 1 p.m. receives over a quarter of its mails in this timeframe. Test account #4 shows a similar result for its reading time at 10 p.m. In view of the magnitude of the effect, Figure 9.5 provides strong evidence in favor of a systematic variation in the communication style of the e-mail sender. On the other hand, it has to be noted that only a single sender in our sample adapts sending times to users' reading time preferences.

9.5 Discussion

The analysis of the e-mails gathered during the 10-week period shows that over 92% of the e-mails from the UK, US and Germany contain tracking images. The prevalence of tracking within the online retail area supports previous studies on the wide application of e-mail tracking mechanisms. During the newsletter signup, companies typically present their privacy terms and conditions. We investigated the statements for the newsletters used in this study. Only 21 out of 52 mention the possibility of using the data gathered for personalization and individualization. Especially newsletters from the UK and US-firms use tracking images to target customers individually while not stating this in their privacy statements. On the other hand, German companies that employ tracking images consistently state this explicitly in their privacy statements. Some German newsletters also offer the option to choose whether companies may use the gathered information for personalization. Given the ubiquity of e-mail tracking and the lack of transparency regarding its use, this study extends the literature by providing an analysis on the reaction of trackers to observed recipient behavior.

Overall, we observe adjustments in sending behavior for 13 out of 44 marketing newsletters, with adjustments affecting the sending behavior. We find that senders respond most often to e-mail opening actions or the lack thereof. Several companies adjust their sending behavior upon realizing that receivers do not open newsletters. Considering users' reading time patterns, we find evidence that a single company within the sample adjusts their communication to accommodate the simulated reading times. We find no evidence that opening a mail multiple times or the location of e-mail access impacts sending behavior.

Surprisingly, we are unable to identify systematic and statistically significant personalization of e-mail content based on information collected through e-mail tracking. While we suspect personalization of product offerings based on the simulated device type for one company, further research is necessary to confirm our findings. While targeting customer individually according to their preferences and interests can be expected to increase click-through and product sales (Golrezaei et al., 2014), content personalization seems to be based on other data, such as previous purchases or similar user interaction.

This study exhibits some limitations that give rise to future research. First, the ten-week time frame for data collection assumes companies to react relatively swiftly. Although some of the newsletter senders adjusted their behavior in this time span, especially companies with a less frequent newsletter delivery may have not had enough user data to systematically react to the user. A longer observation period may be necessary to identify adjustments for non-frequent newsletter.

Another limitation, which simultaneously is the explicit focus of this study, is the concentration on tracking images. Since the data gathered through tracking images needs to be analyzed and interpreted, it could be more complex to employ this as a basis for individualization than it might be to use other techniques (e.g., tracking links). To enhance the results and ideas of this study, it would be useful to include further types tracking mechanisms in further analyses. To consider tracking links in a follow-up study would add another dimension to the behavioral aspects. Another aspect would be selective reading of mails (e.g., select which mail to open based on their title).

The simulated user behavior was restricted to reading behavior of the e-mail. In practice, it is likely that user behavior on the website of the company, matched via the user account or link tracking, provides additional data to be used for personalization of marketing messages. Simulation of actual browsing and purchasing behavior, while complex to conduct in an automated fashion, has the potential to uncover additional personalization strategies, e.g., retargeting of abandoned products, which could be considered in further studies.

Furthermore, we cannot entirely rule out the possibility that some senders might have recognized our test accounts as artificial. The software components cause the simulated users to behave in a very consistent manner, which is unlikely for real e-mail users. Some newsletter senders might have realized this unusual pattern and might have reacted to it. On the other hand, use-cases for fake e-mail accounts that only read e-mails seem rather limited (e.g., compared

to fake accounts for spamming, which would send a huge amount e-mails). In this regard, it is questionable whether senders have implemented sophisticated detection strategies for this kind of suspicious behavior we rely on in this study. For future studies, it would be useful to integrate more random behavior into the experiment to prevent detection mechanisms from uncovering our mail-reading engine.

Finally, a statistical limitation comes from the fact that we have only twelve accounts available. This comes from the vast effort to manually subscribe, for each account, to a large number of newsletters. However, empirical evidence in the form of descriptives and mean comparison across groups clearly suffers from the number of accounts, which, although substantially larger than what has been considered in prior work, is relatively small. We argue that this issue is inherent to the research problem and cannot be overcome easily. Captchas prevented an automation of newsletter subscriptions. On the other hand, crowdsourcing supporters of such research to assist with manual labor would inevitable raise awareness of the research, which might carry over to mailers and thus introduce bias. In this regard, we consider the results presented here as a valuable first evidence into a sparsely researched phenomenon but also strongly encourage further research to expand the scale of the analysis.

The study focused on the trading industry, even though many other industries are equally relevant. Future studies should therefore incorporate other industries to gain a more diverse picture on the application of e-mail tracking mechanisms.

9.6 Conclusion

E-mail tracking facilitates gathering information regarding the individual recipient reading behavior. Former studies reveal that professional e-mail newsletters commonly include tracking elements (Fabian et al., 2015). However, former studies do not check whether information that can be gathered through the tracking images is actually used by the e-mail senders.

Our experiment strives to close this gap and validates the usage of e-mail tracking information by e-mail senders. To that end, the study uses twelve mail accounts each of which simulates an individual user behavior to gather newsletters over a 10-week period from retail companies across Germany, the UK, and the USA. We find that 92% of the e-mails from the UK, US and Germany contain tracking images.

The experimental data shows that the tracking images detected are in fact used to assess individual behavior and to adjust marketing communication on the individual level. This confirms the relevancy of the potential threats to user privacy resulting from e-mail tracking. Even though e-mail clients typically allow to block external referenced content, such as images, and thereby counteract e-mail tracking images, studies revealed this blocking approach to be impractical for image-rich e-mails, such as newsletters. Blocking images completely is not assumed to be an effective strategy, which is why selectively blocking content is suggested (Bender et al., 2016). The experiment further reveals that companies employ personal response data to individualize marketing communication for newsletters. We observe individualization in 13 out of 44 (30%)

newsletters. We find statistically significant response of senders to e-mail opening. In reverse, several companies stop delivering newsletters upon realizing that receivers do not open newsletters. With regard to time patterns, we find a single company to adjusts their newsletter mailings to accommodate the different simulated reading times. No evidence was found, that multiple openings as well as location related aspects impact sending behavior. We find a statistically weak significance for device-category specific adjustment of mail frequency, with mobile devices receiving slightly less e-mails than the desktop accounts. With regard to content adjustments, we are unable to show systematic variations as expected for online retailers. Nonetheless, a single company adjusted the products offered to infrastructure characteristics.

Observed adjustment patterns can be considered easy-to-implement options to individualize communication, especially for e-commerce retailers due to their typically advanced IT-systems regarding analytics and wide product portfolio. Targeting users individually is also important for such companies to succeed in customer acquisition, growth, and retention. Regarding legal issues, we find that that not all companies which adapt the mail sending behavior inform subscribers of such mechanisms in their privacy policy. These results confirm that data collection, storage and analysis on the personal level takes place and emphasizes the need for additional research regarding the extent of identified privacy risks and the development of efficient protection strategies.

Bibliography

Acar, G., Eubank, C., Englehardt, S., Juarez, M., Narayanan, A., & Diaz, C. (2014). The Web Never Forgets: Persistent Tracking Mechanisms in the Wild, In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, Scottsdale, Arizona, USA, ACM. https://doi.org/10.1145/2660267.2660347

Agosti, M., & Di Nunzio, G. M. (2007). Gathering and Mining Information from Web Log Files, In *Digital Libraries: Research and Development*, Berlin, Heidelberg, Springer. https://doi.org/10.1007/978-3-540-77088-6_10

Bender, B., Fabian, B., Lessmann, S., & Haupt, J. (2016). E-Mail Tracking: Status Quo and Novel Countermeasures, In *Proceedings of the 37th International Conference on Information Systems (ICIS)*, AIS.

Bouguettaya, A., & Eltoweissy, M. (2003). Privacy on the web: Facts, challenges, and solutions. *IEEE Security & Privacy Magazine*, *1*(6), 40–49. https://doi.org/10.1109/MSECP.2003.1253567

Brookman, J., Rouge, P., Alva, A., & Yeung, C. (2017). Cross-device tracking: Measurement and disclosures. *Proceedings on Privacy Enhancing Technologies*, *2017*(2), 133–148. https://doi.org/10.1515/popets-2017-0020

Bujlow, T., Carela-Espanol, V., Sole-Pareta, J., & Barlet-Ros, P. (2017). A survey on web tracking: Mechanisms, implications, and defenses. *Proceedings of the IEEE*, *105*(8), 1476–1510. https://doi.org/10.1109/JPROC.2016.2637878

Cselle, G., Albrecht, K., & Wattenhofer, R. (2007). BuzzTrack: Topic Detection and Tracking in Email, In *Proceedings of the 12th International Conference on Intelligent User Interfaces*, Honolulu, Hawaii, USA, ACM Press. https://doi.org/10.1145/1216295.1216331

EHI Retail Institute. (2019). B2c-e-commerce: Ranking der top100 grösss ten online-shops nach umsatz in deutschland im jahr 2018.

Englehardt, S., & Narayanan, A. (2016). Online Tracking: A 1-million-site Measurement and Analysis, In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Vienna, Austria, ACM. https://doi.org/10.1145/2976749.2978313

Ermakova, T., Fabian, B., Bender, B., & Klimek, K. (2018). Web Tracking: A Literature Review on the State of Research, In *51nd Hawaii Conference on System Sciences*. https://doi.org/10.24251/HICSS.2018.596

Ermakova, T., Hohensee, A., Orlamünde, I., & Fabian, B. (2017). Privacy-invading mechanisms in e-commerce: A case study on German tourism websites. *International Journal of Networking and Virtual Organisations*, *20*(2), 105–126. https://doi.org/10.1504/IJNVO.2019.097629

Evans, M., & Furnell, S. (2003). A model for monitoring and migrating web resources. *Campus-Wide Information Systems*, *20*(2), 67–74. https://doi.org/10.1108/10650740310467763

Fabian, B., Bender, B., & Weimann, L. (2015). E-Mail Tracking in Online Marketing: Methods, Detection, and Usage, In *12th International Conference on Wirtschaftsinformatik*, Osnabrück, Germany.

Falahrastegar, M., Haddadi, H., Uhlig, S., & Mortier, R. (2016). Tracking Personal Identifiers Across the Web, In *Passive and Active Measurement (PAM '16)*, Cham, Springer International Publishing. https://doi.org/10.1007/978-3-319-30505-9_3

Foulger, M. G., Chipperfield, T. R., Cooper, J. S., & Storms, A. C. (2008). *System and Method Related to Generating and Tracking an Email Campaign* (US20080244027A1).

Golrezaei, N., Nazerzadeh, H., & Rusmevichientong, P. (2014). Real-time optimization of personalized assortments. *Management Science*, *60*(6), 1532–1551. https://doi.org/10.1287/mnsc.2014.1939

Gomer, R. C., Mendes Rodrigues, E., Milic-Frayling, N., & Schraefel, M. (2013). Network Analysis of Third Party Tracking: User Exposure to Tracking Cookies through Search, In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT '13)*.

Hamed, A., Kaffel-Ben Ayed, H., Kaafar, M. A., & Kharraz, A. (2013). Evaluation of Third Party Tracking on the Web, In *8th International Conference for Internet Technology and Secured Transactions (ICITST-2013)*. https://doi.org/10.1109/ICITST.2013.6750244

Han, S., Jung, J., & Wetherall, D. (2012). *A Study of Third-Party Tracking by Mobile Apps in the Wild* (Technical Report UW-CSE-12-03-01). University of Washington.

Jin, L., Takabi, H., & Joshi, J. B. (2010). Security and Privacy Risks of Using E-mail Address as an Identity, In *Proceedings of the Second International Conference on Social Computing*, IEEE. https://doi.org/10.1109/SocialCom.2010.134

Libert, T. (2015). Exposing the invisible web: An analysis of third-party HTTP requests on 1 million websites. *International Journal of Communication*, *9*, 3544–3561.

Mayer, J. R., & Mitchell, J. C. (2012). Third-Party Web Tracking: Policy and Technology, In *IEEE Symposium on Security and Privacy*, San Francisco, CA, USA, IEEE. https://doi.org/10.1109/SP.2012.47

Mitchell, I. D. (2012). *Third-Party Tracking Cookies and Data Privacy* (SSRN Scholarly Paper). Rochester, NY.

Moscato, D. R., Altschuller, S., & Moscato, E. D. (2013). Privacy policies on global banks' websites: Does culture matter? *Communications of the IIMA*, *13*(4), 91–109.

Moscato, D. R., & Moscato, E. D. (2009). Information security awareness in e-commerce activities of B-to-C travel industry companies. *International Journal of the Academic Business World*, *3*(2), 39–46.

Nikiforakis, N., Maggi, F., Stringhini, G., Rafique, M. Z., Joosen, W., Kruegel, C., Piessens, F., Vigna, G., & Zanero, S. (2014). Stranger Danger: Exploring the Ecosystem of Ad-based URL Shortening Services, In *Proceedings of the 23rd International Conference on World Wide Web*, Seoul, Korea, ACM. https://doi.org/10.1145/2566486.2567983

O'Connell, J. (2014). The State Of Programmatic Media.

Oppliger, R. (2007). Providing certified mail services on the internet. *IEEE Security and Privacy Magazine*, *5*(1), 16–22. https://doi.org/10.1109/MSP.2007.15

Parra-Arnau, J. (2017). Pay-per-tracking: A collaborative masking model for web browsing. *Information Sciences*, *385-386*, 96–124. https://doi.org/10.1016/j.ins.2016.12.036

Roesner, F., Kohno, T., & Wetherall, D. (2012). Detecting and Defending Against Third-Party Tracking on the Web, In *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*, San Jose, CA, USENIX Association.

Sanchez-Rola, I., Ugarte-Pedrero, X., Santos, I., & Bringas, P. G. (2017). The web is watching you: A comprehensive review of web-tracking techniques and countermeasures. *Logic Journal of IGPL*, *25*(1), 18–29. https://doi.org/10.1093/jigpal/jzw041

Schelter, S., & Kunegis, J. (2016a). On the Ubiquity of Web Tracking: Insights from a Billion-Page Web Crawl. *arXiv preprint*, arXiv:1607.07403 [cs].

Schelter, S., & Kunegis, J. (2016b). Tracking the Trackers: A Large-Scale Analysis of Embedded Web Trackers, In *Tenth International AAAI Conference on Web and Social Media*.

Schmidt, J. (2013). E-Mail im Visier: Tracking im Alltag aufspüren und abstellen. *c't Magazin für Computertechnik*, *22/2013*, 130–135.

Suneetha, K., & Krishnamoorthi, R. (2009). Identifying user behavior by analyzing web server access log file. *International Journal of Computer Science and Network Security*, *9*(4), 327–332.

Thode, W., Griesbaum, J., & Mandl, T. (2015). "I Would Have Never Allowed It": User Perception of Third-party Tracking and Implications for Display Advertising, In *Proceedings of the 14th International Symposium on Information Science (ISI '15)*. https://doi.org/10.5281/zenodo.17971

UK, I. R. M. (2016). The 2015 IRUK Top500 by Performance Cluster.

Zaczkiewicz, A. (2016). Amazon, Wal-Mart Lead Top 25 E-Commerce Retail List. *WWD*.

Chapter 10

Robust Identification of Email Tracking: A Machine Learning Approach

Publication
Haupt, J., Bender, B., Fabian, B., & Lessmann, S. (2018). Robust identification of email tracking: A machine learning approach. European Journal of Operational Research, 271(1), 341–356. https://doi.org/10.1016/j.ejor.2018.05.018

Abstract
Email tracking allows email senders to collect fine-grained behavior and location data on email recipients, who are uniquely identifiable via their email address. Such tracking invades user privacy in that email tracking techniques gather data without user consent or awareness. Striving to increase privacy in email communication, this paper develops a detection engine to be the core of a selective tracking blocking mechanism in the form of three contributions. First, a large collection of email newsletters is analyzed to show the wide usage of tracking over different countries, industries and time. Second, we propose a set of features geared towards the identification of tracking images under real-world conditions. Novel features are devised to be computationally feasible and efficient, generalizable and resilient towards changes in tracking infrastructure. Third, we test the predictive power of these features in a benchmarking experiment using a selection of state-of-the-art classifiers to clarify the effectiveness of model-based tracking identification. We evaluate the expected accuracy of the approach on out-of-sample data, over increasing periods of time, and when faced with unknown senders.

10.1 Introduction

Data on email reading behavior is routinely used to infer commercially valuable information from customers. For example, it allows marketers to derive user profiles and measure the reach and effectiveness of email marketing campaigns (Hasouneh & Alqeed, 2010). It also facilitates marketing activities, such as calling prospective customers at the time they open a marketing message (Hlatky, 2013). The Direct Marketing Association estimates that its members achieved an average return of £38 for every pound spent on email marketing and that this return on investment will continue to increase in the future with the spread of advanced testing and personalization (The Direct Marketing Association, 2015). This gives marketers a strong incentive to monitor how customers interact with email newsletters and advertising. Using the same methods, spammers and phishers rely on email tracking to validate and collect

active email addresses for their illegal activities (Vaas & Stockley, 2014). Current email tracking techniques enable the sender to track if and how often an email is opened, the time at which the email is read, which device as well as operating system the recipient uses, and her Internet Protocol (IP) address (Murphy, 2014). Such information, in turn, facilitates deducting the location of the reader, her affiliation to a company or organization, email reading behavior, travel patterns based on desktop and mobile use, and if an email was forwarded or printed (Technology Analysis Branch, 2013). A peculiarity of email tracking is that tracking information is linked to a user's email address, which is an almost unique identifier of the user that can easily be matched to other accounts of the user such as social media profiles. Consequently, tracking users across devices, applications, locations, etc. is much easier in email tracking compared to other channels such as web tracking. Importantly this data is typically gathered without active consent, case-by-case confirmation or even awareness of the recipient. In combination, these characteristics facilitate surveillance and constitute an invasion of user privacy. As we are able to show, email tracking does not merely constitute a theoretical risk but is ubiquitous in marketing communication.

Therefore, email users require tools to protect against potential privacy hazards caused by email tracking. A review of the literature and contemporary email clients reveals a lack of easy-to-use, effective, and reliable protection methods. The reason is that the identification of tracking images, which are the main tracking mechanisms in emails, poses specific challenges that render standard ad blockers and blacklists ineffective. The goal of this paper is to contribute towards empowering email users to protect their privacy. To that end, we develop a machine learning approach to detect tracking elements in emails with the ultimate goal to filter them selectively.

The contribution of this paper is three-fold. First, we establish the prevalence of email tracking through the analysis of 30,756 marketing-communication emails from 300 global companies collected over a period of 20 months. We extend previous analyses by comparing the occurrence of email tracking in different industries and identifying common email-tracking providers. Second, we develop a set of features geared towards the identification of tracking images under real-world conditions. These features are devised to be computationally efficient, to generalize to structures of unseen tracking images, and to be resilient against changes in tracking structures over time. Third, using a selection of state-of-the-art classifiers, we test the predictive power of these features in a benchmarking experiment to clarify the effectiveness of model-based tracking identification. We evaluate the expected accuracy of the approach on test sets that are out-of-sample, out-of-time, i.e. after increasing amounts of time have passed, and out-of-universe, i.e. when faced with unknown senders. This allows us to identify an optimal identification model and appraise the degree to which a model-based approach protects against email tracking in application.

The remainder of the paper is structured as follows. Section 10.2 introduces current email tracking techniques. Section 10.3 identifies related literature. Section 10.4 examines the occurrence of tracking within the commercial newsletters that we collect for the study to stress the relevancy of defensive strategies. Section 10.5 presents the featurization methodology to

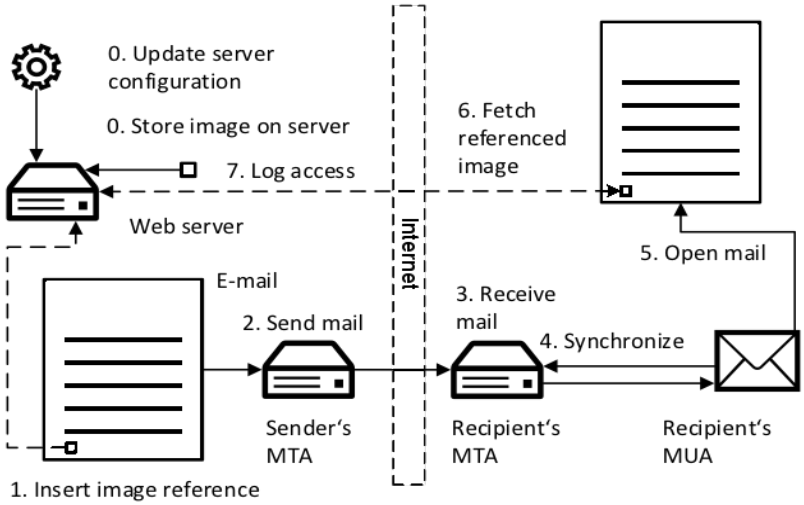


Figure 10.1: Overview of the email tracking system and process

identify tracking images. Section 10.6 and Section 10.7 elaborate on the experimental design and empirical results, respectively. Section 10.8 concludes.

10.2 E-Mail Tracking Technology

We start by outlining email tracking methodology and the degree to which it impacts user privacy. This section provides the technical foundation to develop features for tracking identification and countermeasure design. The tracking process (Figure 10.1) is based on emails that are written in Hypertext Markup Language (HTML) referencing specific external resources. Prior literature refers to these resources with different terms, including "web bugs" (Martin et al., 2003) and "tracking pixels" (Vaynblat et al., 2009). Considering their function and location within the email's HTML code as <img> tags, we use the term *tracking images*. The tracking process starts with the sender dispatching an HTML-based email. The email includes an image tag, which references a tracking object stored on a server of the sender, or its tracking provider, in the form of a Uniform Resource Locator (URL). When the recipient opens her mail client, the mail user agent (MUA) synchronizes the local mail repository with updates provided by the recipient's message transfer agent (MTA), and the user receives the email. When the recipient opens the email, which contains the tracking image tag, the mail client requests the referenced file. The web server, where the file is stored, logs this request and provides the image to the client. Log analysis allows the sender to infer information on the recipient's access device and email reading behavior. For example, if the email is opened on different devices, every individual access is logged with the corresponding user-agent information, which allows for cross-device tracking.

Image requests themselves do not contain sufficient information to identify a specific email recipient. For the purpose of matching an image request to a known recipient and thus track individual behavior, either the tracking object must be unique to the recipient or the reference URL must contain a unique tag that identifies the recipient. In both cases, the reference within

the <img> tag will be unique to the email recipient. By sending images with a specific reference to only one recipient, trackers control that subsequent access to the image via that reference can be attributed to a single recipient. The hash of the recipient's email address has been identified as a common approach to create anonymized identifiers (Englehardt et al., 2018).

In contrast, requests to non-tracking images from references that are common to all recipients, e.g. product pictures, are logged on the server in combination with the respective IP address and device information but cannot be linked to recipients' email addresses. An extension to this form of aggregate data collection are images containing, for example, an identifier for the email campaign rather than individual recipients. Use cases of tracking on the aggregated level (i.e., without an identifier for individual users) include measuring the opening rate of an email campaign for A/B testing of newsletter design. Since no individual information is collected by non-unique tracking images, their privacy implications are less pronounced. We consequently focus on images that include a unique identifier and facilitate tracking of individual users in this study. For readability, we refer to individual tracking images as tracking images.

Individual tracking data poses a privacy risk because personal information about the identity and behavior of the tracked user can be derived without her consent or awareness. The log entries facilitate inducing that the user has read or at least opened an email, because current email clients do not download images before the corresponding email is opened. In case of spam emails sent to random email addresses, this is sufficient to prove that an active account has been found. In addition, the time stamp and the existence of multiple log entries reveal the time of day and the number of times an email is opened. The combination of multiple entries for a single mail, as well as multiple entries from one user for different mails, provide insights into the recipient's email reading behavior. Furthermore, the log entry facilitates inferring information about the user environment (Agosti & Di Nunzio, 2007). Data on the use of mobile or desktop devices, especially when aggregated over time, conveys additional information about user activities such as office or travel times. It is also possible to track whether an email has been printed through a print stylesheet, either by tracking the stylesheet directly or by matching stylesheet access to the device information collected by the tracking image.

More complex analysis reveals additional information about the recipient. For example, transmitted IP addresses enable trackers to gather location-related information (Poese et al., 2011). Based on a reverse lookup of an IP address, a log entry may also reveal a user's affiliation to an organization, for example, if private emails are opened at work. Combining pieces of information also facilitates predicting whether an email has been forwarded and allows deducting travel routines. For example, a major technology company combined the IP address, location information and the time stamp of a log entry to identify a board member who was forwarding confidential information (Evers, 2006).

A crucial point differentiating web and email tracking is that the collected and combined information is not anonymous in email tracking. While both rely on similar mechanisms (e.g., cookies or tracking images) and gathers a rich set of behavioral information, users tracked via web tracking are not directly personally identifiable without consent. The personal identifica-

tion of the tracked user is often impossible and alternative methods to recognize users over time and web sites have been proposed (Nikiforakis et al., 2013; Yang, 2010). Information collected via email tracking, on the other hand, is necessarily linked to an email address, which provides a platform independent almost unique identifier of a person and often contains the user's name and possibly organization. Additionally, it is often possible to link an email address to personal online profiles, for example on social media sites.

Currently, the only solution for providing fully reliable privacy protection against email tracking in HTML emails is to block all external content referenced in emails. From a technical point of view, this approach is easy to implement on either the server or the client and can be activated as default for most email clients. However, blocking all images in an email entails a substantial loss of information and interferes with user experience by excluding all referenced images and the corresponding content. Possible further issues include incorrect formatting, loss of styling elements, and misinterpretation if external images convey crucial information.

A selective filtering approach provides a balance between preventing user tracking and sustaining user experience. It operates through identifying and selectively blocking tracking elements within an email. In this approach, a predictive model is used to categorize referenced images into tracking and non-tracking images. Non-tracking images remain untouched, whereas tracking image references are removed from the email. Note that tracking images are often transparent and do not contain content (Bender et al., 2016). In the ideal case, the user avoids being tracking without noticing that an email has been sanitized. However, the efficacy of selective filtering depends critically on the algorithm for tracking image identification.

10.3 Related Work

We organize the literature related to this study into three categories. First, we summarize the existing research on email tracking. Given the sparsity of research on this specific topic, we next identify studies on web tracking, which is similar from a technological perspective. Last, we discuss previous studies investigating mechanisms to selectively remove unwanted elements from HTML-based content.

Email tracking is periodically covered in the general press, where it is criticized for invading privacy (Murphy, 2014) or mentioned as a tool to uncover information leakage (Hodgekiss, 2010). Some authors hint at the possibility of tracking in HTML emails (Bouguettaya & Eltoweissy, 2003; Harding et al., 2001; Martin et al., 2003; Moscato et al., 2013). Few academic papers have examined the topic. A notable exception is the recent study by Englehardt et al. (2018) showing the ubiquity of email tracking in a large scale sample. Most studies focus on marketing rather than privacy or countermeasures against email tracking, for example Bonfrer and Drèze (2009) and Hasouneh and Alqeed (2010), who structure technical and process-related aspects of email tracking from a marketing perspective and stress the importance and prevalence of tracking in newsletters and other marketing communication. This study extends our own previous research on the characteristics of email tracking images as well as mechanisms for tracking detection and prevention (Bender et al., 2016) in three ways. First, we broaden the

scope of the analysis of tracking prevalence through examining emails gathered over a horizon of 20 months and from 33 industries. Second, we substantially improve the tracking detection engine. Whereas Bender et al. (2016) use an untuned feed-forward neural network classifier, we conduct a comprehensive benchmark of state-of-the-art machine learning algorithms for tracking image classification. Third, we propose novel predictors of email tracking to ensure deployability. Most importantly, we establish the stability of detection accuracy and generality of the tracking protection framework through rigorous out-of-time and out-of-universe testing. This allows us to demonstrate that the proposed system is adequate to protect users against privacy invasions under real-world conditions.

From a technological point of view, email tracking can be considered an adaptation of web tracking mechanisms to HTML-based emails. Unlike email tracking, the use of web tracking in different situations (Javed, 2013; Jensen et al., 2007) and its detection (Alsaid & Martin, 2003; Fonseca et al., 2005) have received much attention in the literature. Prevention of such mechanisms and the evaluation of existing software solutions have also been studied (Fonseca et al., 2005; Leon et al., 2012). Other research emphasizes the technical aspects of web tracking, such as different categories of web bugs (Dobias, 2011) or the potential for aggregating multiple server log files (Evans & Furnell, 2003). We make use of the mature research towards the detection of web tracking and extend it to email tracking.

Methodologically, the identification of tracking content is related to the identification of tracking and advertising on web pages (e.g., ad blocking) or other unwanted content in emails (e.g., spam and phishing detection). These applications make use of information related to the image reference URL, the email sender, the website host, the content visible to the user, and the formatting of an image. Ad blockers rely on the image content for classification (Li et al., 2010). However, content classification requires accessing the image, which would be registered by the tracking server. Therefore, content-based approaches are inapplicable to prevent email tracking effectively.

An alternative is to examine the structure of image references. Li et al. (2010) and Kushmerick (1999) propose a range of features to identify advertising images on web pages. They focus on the formatting and image-reference link relative to other images on the same page; for example, investigating whether the image domain is different from the site domain, with a deviation being indicative of third-party content. The reference structure itself is also used to identify advertisement. For example, Shih and Karger (2004) propose a heuristic that exploits the fact that advertisement images are often placed in a different folder than content images. URLs have also been used with success to identify phishing mails (Blum et al., 2010; Garera et al., 2007; Ma et al., 2009b; Whittaker et al., 2010) and to classify web pages (Kan & Thi, 2005; Shih & Karger, 2004).

Most of the above approaches rely on identifying keywords through text mining on parts of the URL. These keywords include both words in natural language describing the target-link content, meaningful letter or number combinations called tokens, and recurring server or folder names. While these can be identified for tracking images, they require constant updating and

are susceptible to avoidance strategies by spammers and trackers, respectively. Fette et al. (2007) introduce predictors counting the number of dots and the number of different top-level domains in mail links to capture the complexity of the URL and the increasing number of domains involved in phishing. We extend these ideas when creating features to capture the structure of tracking image references.

Especially for phishing analysis, some approaches rely on the content of the email. Bergholz et al. (2008) propose features based on a dynamic Markov chain and topic models based on Latent Dirichlet Allocation. We focus on tracking image identification but acknowledge that a pre-classification of emails based on their subject line or content could convey some preliminary information on the probability of an email being tracked. Preliminary classification could increase speed and accuracy of tracking identification in future work.

Host information has been found effective in phishing and spam detection (Fette et al., 2007; Ma et al., 2009a, 2009b). This information is gathered via the IP address and a WHOIS request to the domain of the server that hosts a referenced website, because a phishing site "may be hosted in less reputable hosting centers, on machines that are not conventional web hosts, or through disreputable registrars" (Ma et al., 2009b). This reasoning does not hold for email tracking in e-commerce, where businesses operate within legal bounds and tracking images are hosted on official company or contractor servers. Moreover, looking up external information slows down the identification process in potential real-time applications (Blum et al., 2010).

In summary, prior work in the context of web tracking mentions the existence of email tracking, hints at tracking methods, and criticizes privacy implications. However, we find a lack of research investigating the prevalence of (legal) tracking activities and approaches to email prevent tracking. While there exist initiatives to develop anti-tracking software in the form of modified mail clients and add-ons that support selective tracking prevention (Barrett, 2015), these tools are unable to provide reliable protection against most tracking approaches (Bender et al., 2016). Therefore, we extend prior work through studying a more comprehensive set of data and providing the foundation of a detection system to identify and selectively block tracking images. Specifically, we build on existing predictors of tracking use and extend these so as to ensure feasibility and improve resilience in real-world applications. Our empirical analysis then establishes the best learning algorithm for the task of tracking image detection and estimates its performance on emails from senders not seen in the data, and also after periods of time between training and application.

10.4 Data and E-Mail Tracking Usage

Analyzing the occurrence of tracking and training a supervised learner for automated detection require data on email communication including the status (tracking/non-tracking) of every image across all emails in the data. The collection of this ground truth data is complex due to important differences in data collection between ad blocking or spam detection and the identification of tracking images. In particular, comparable studies obtain status labels through human judgment, which is often crowd-sourced. The information available for classification in

Table 10.1: Example image tags of two tracking and non-tracking images, respectively. Tracking image tags are shown in rows 2 and 3.

(1) <img border="0" alt="" src="https://i.emlfiles.com/cmpimg/[...]/w37_33215_t.jpg>
(2) <img src="http://[company].msgfocus.com/t/1nqheZESyay9K7.png" alt="">
(3) <img src="http://newsletter.[company].de/tr/p.gif?uid=af[...]&mid=3f[...]" width="1" height="1" [...]
(4) <img src="https://img.srv2.de/bm/img/c4/a/c4[...].png" width="8" height="6" [...] border="0">

our setting consists of the images themselves and the email source code. Identification of tracking images based on the image content is unreliable, since content and tracking functionality are independent of one another. In practice, transparent or tiny images without actual content are also legitimately used for formatting purposes (Martin et al., 2003). Thus, ground truth classification must be based on the image tag in the email code and, most importantly, the image reference. Image references do not have to be human-understandable, and tracking images are hidden from the recipient by design, which makes identification through human judges unreliable; as illustrated in Table 10.1.

A constituent property of personal tracking image references is that they contain a unique identifier for the recipient (see Section 10.2). We therefore create two identities and corresponding email addresses using Gmail and match the emails and images received on both accounts to identify tracking elements. We do this by extracting images from the HTML content of each pair of emails sent to both accounts and comparing the image reference URLs at each position for differences. Images for which the reference URLs are an exact match are classified as non-tracking images and images with different URLs as tracking images. To avoid bias from senders changing their email policy in response to the reading behavior of the users they are tracking, we ensure that none of the external images are requested from the web server at any point.

With each account, we signed up for the newsletters of 300 companies and collected emails in a 20-month period from 2015 to 2017. Although not representative of email communication in general, we argue that newsletter emails are a suitable vehicle for this analysis. First, it is likely that companies use email tracking to assess the effectiveness of their newsletters (Hasouneh & Alqeed, 2010). We aim to increase this likelihood by concentrating on large companies, which are on average faster to adopt novel technology (Premkumar & Roberts, 1999). We sign up to email newsletters from the top-100 companies ranked by revenue in Germany, Great Britain, and the United States. Second, the wide availability of different newsletters simplifies systematic data collection and facilitates the gathering of a large amount of data. At the same time, signing up to newsletters requires an active request restricting the amount of data and its variance and may introduce selection bias, as opposed to, for example, the passive collection of unsolicited emails in spam detection. To mitigate this effect and ensure substantial variance, our company selection is based on company size and includes companies from three countries. Third, newsletter can be ordered multiple times without difficulty. In contrast to for example personal communication, using commercial newsletters is an effective way to gather ground-truth data.

Each artificial identity received 30,756 emails, which we could match between accounts. Of

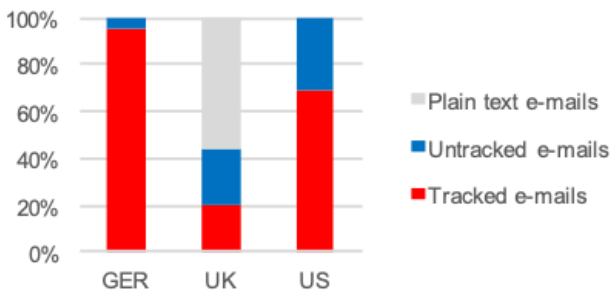


Figure 10.2: Ratio of tracked emails per country

these, 7,154 (23%) are in plain-text format, while the remaining 23,602 are HTML-based and thus facilitate tracking. Of the HTML emails, 21,500 (91%) contain a total of 794,519 external image references, which constitute the data set on which we build and test the tracking detection model. The number of images per email varies considerably and shows positive skewness. We observe a mean (median) value of 37 (18) external images per email. 16,410 emails (69% of HTML emails) contain tracking elements, which illustrates that tracking is common in company newsletters. The ratio of emails received from each country roughly corresponds to the ratio of companies with 29% of emails sent by companies from Germany, 40% from the United Kingdom (UK), and 31% from the United States (US). The tracking quota and the fraction of HTML-based emails vary significantly between countries (see Figure 10.2). The ratio of HTML emails is close to 100% for Germany and the US. In the UK, only 44% of emails are in HTML format, and out of these, only 46% are tracking mails, resulting in an overall tracking quota of 20%. This is significantly lower than the tracking quotas in Germany (95%) or the US (69%).

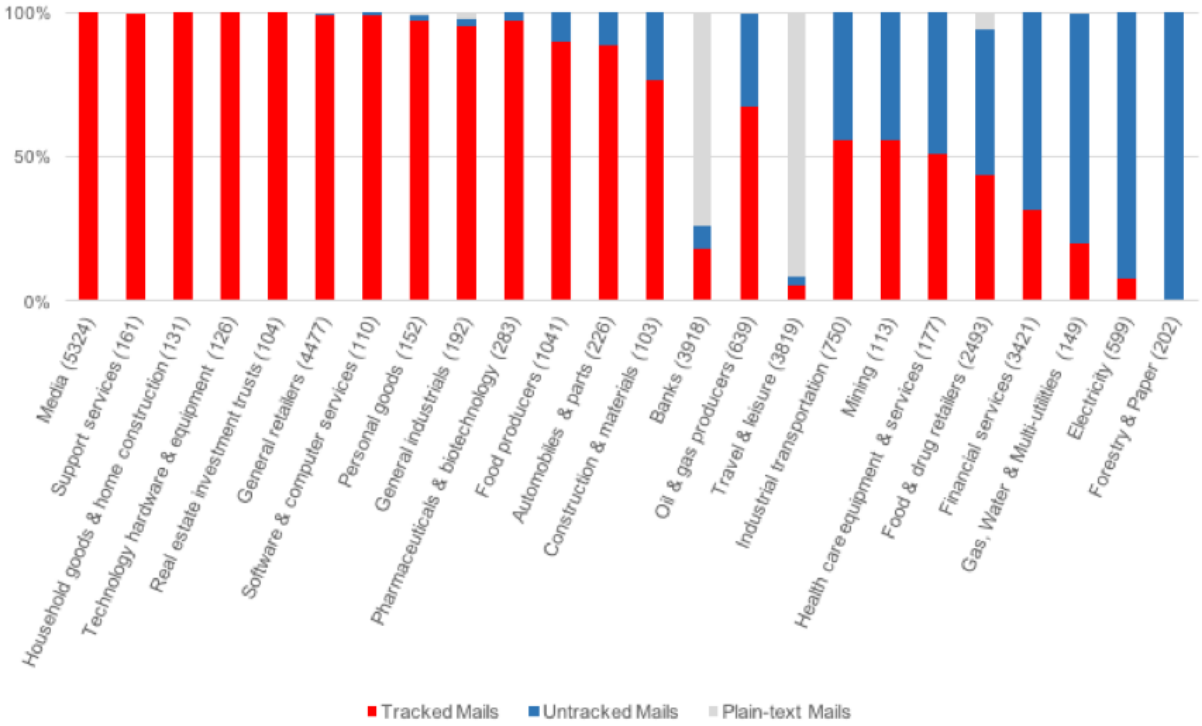


Figure 10.3: Ratio of tracking by industry (total number of emails in brackets)

Country-level variation reflects the industry distribution of the top companies in each of the

three countries. Each email is matched to a company according to its sender domain and assigned to an industry category based on the Financial Times Equities database (Financial Times, 2017). Figure 10.3 presents the per-industry tracking ratio for industries with more than 100 emails in the sample. We observe that customer-targeted newsletters are tracked with near certainty, while business-to-business newsletters and company news, predominant among industrial producers, are less likely to contain a tracking image. An exception to this rule are investor bulletins, which are sent at high frequency in plain text. Bulletins are responsible for the large fraction of plain-text emails (light grey color) observed for the banking and travel sector.

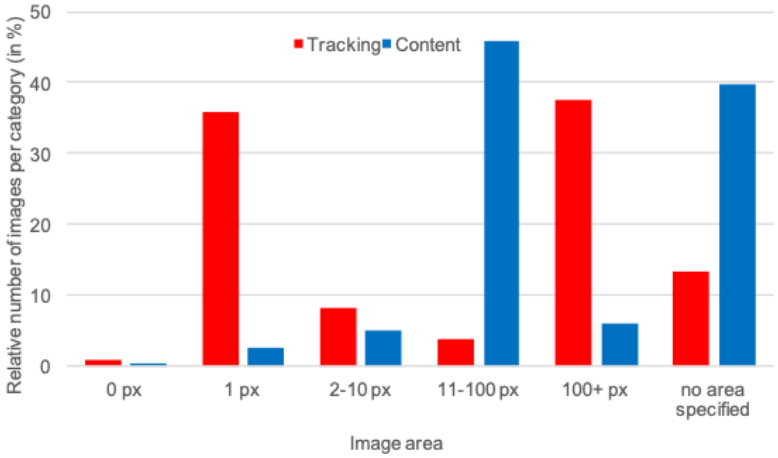


Figure 10.4: Image area (height×width) for tracking and content images

The tracking literature assumes tracking images to be small, typically with an area of 1 square pixel (Martin et al., 2003). We analyze the observed image sizes for tracking and content images in Figure 10.4. 35% of the tracking images for which a size could be determined have an area of one square pixel. There exist images with a specified area of 0, which are most likely not shown by the email client thus making them effectively invisible. The majority of tracking images has an area above 100 square pixels (38%) or no specified size (13%). Note that we consider several ways to specify the size of images (see Section 10.5.1). The results suggest that simple rules to filter images based on their area are likely to fail.

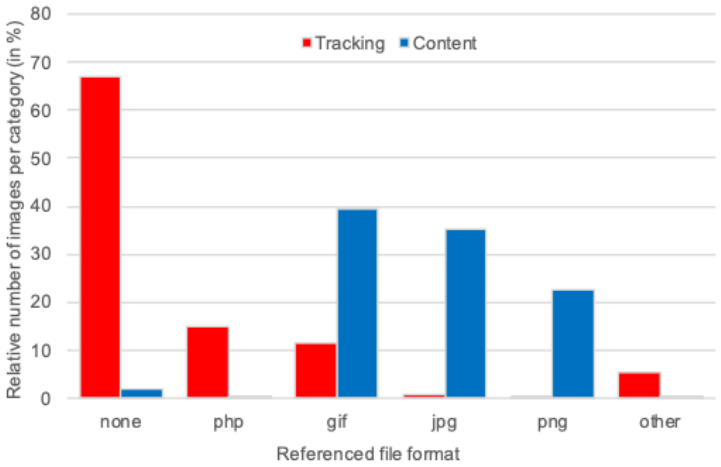


Figure 10.5: Relative frequency of file formats for tracking and non-tracking images

The file extensions extracted from each image reference (Figure 10.5) reveal that the file format is not indicated for two-thirds of tracking files. Approximately 20% of references including a file format indicate the file to be a code script rather than an actual image file, with the majority of scripts written in PHP or ColdFusion Markup. The use of executable files instead of images sheds light on the underlying tracking infrastructure and suggests that the file access and the information associated with it can be dynamically processed or forwarded to internal or third-party databases. The findings suggest the file type – when available – is highly discriminatory for the identification of tracking images.

10.5 Tracking Image Detection

A selective tracking prevention system that targets and filters tracking elements conceptually consists of three components. First, a data-input component extracts all image tags and their attributes from raw email code. Second, a tracking detection engine, which represents the core of the system, performs two tasks. It creates indicative features from the raw data (e.g., HTML image tag) and uses the features as input to a classification model. The model estimates the probability that an image is used for tracking. Third, the selective filtering component processes the estimated tracking probability to handle external images. Rather than blocking all images in an email, which is currently the most secure way to avoid email tracking, the system is able to selectively block the download of images with high tracking probability. This empowers users to see uncritical content without being tracked and to decide, after inspecting a sanitized version of an email, whether they want to permit the download of further, system-filtered images, despite the risk of being tracked. This way, the envisioned system also offers a viable approach to handle content images that perform tracking. More specifically, users are enabled to make a conscious decision how they trade-off the risk of being tracked by the sender of an email with possible readability issues caused by image filtering.

The detection engine classifies unknown images into two categories, *tracking* and *non-tracking/ content* on the basis of meta-data extracted from HTML code. Images used for individual tracking exhibit structural peculiarities, which facilitate such classification. In particular, they contain a unique user identifier assigned by the tracker, are distinctly formatted, and are often handled by a different department or company (Bender et al., 2016). However, correct identification is challenging – even for human judges – for two reasons. First, tracking images do not necessarily fulfill all criteria simultaneously and show significant variation in the observed patterns. While certain structures are necessary or common, their actual format depends largely on choices made by the tracker. For example, the existence of an individual identifier is necessary, but the identifier itself may be constructed out of numbers, lowercase and uppercase letters or any combination thereof and its position in the reference can be as folder name, image name or URL parameter (see Table 10.1). Note that the <img> attribute itself is not transmitted within the request to the webserver and thus not suitable for user tracking. The identification of tracking images is further complicated by the possibility to track images of any format or size, including branding or content images. Second, non-tracking images may display the above characteristics, including very small images used for formatting, or images handled through

content-management systems, whose file names resemble user IDs. Under these restrictions, only complex rules can ensure satisfactory identification of unwanted images at a low rate of false identifications without interfering with the email content or formatting. We therefore employ machine-learning techniques to develop a detection model.

Table 10.2: Predictors for the detection of tracking images by category

Reference structure	**Reference str. (cont.)**	**Email header**
Count IDs in filename*	Reference includes '?'	*Custom header fields*
Count IDs in path*	*Reference includes '@'*	Image name matches sender
Count number strings*	*Reference includes 'id'*	Length 'unsubscribe' field*
Count number-letter changes*	*Reference includes 'click'**	Ref. parts match 'list-unsubscribe'*
Count numbers*	*Reference includes 'open'**	Ref. parts match 'received-spf'*
Count punctuation*	*Reference includes 'track'**	
Count strings	*Reference includes 'view'**	**Image server**
Count uppercase*		Images sharing same domain
Fileformat 'jp(e)g' *	**Image structure**	Matching image and sender domain *
Fileformat 'php'*	*align**	
Fileformat none*	*Area**	
Img. sharing same fileformat	*Area: 0 pixels^2 **	
Filename length	*Area: 1 pixels^2 **	
Image link similarity (Max.)	*Area: 100+ pixels^2 **	
Image link similarity (Mean)	*Area: 11-100 pixels^2 **	
Image link similarity (Min.)	*Area: None specified**	
Image reference (ref.) length*	*Border width**	
Rel. reference length	*Length of attribute 'class'**	
Length of domain	*Contains 'style: display'**	
Longest Number in reference	Count other identical images*	
Difference to mean letter count 'b'	*Image width**	
Difference to mean letter count 'f'	*Length of tag 'Title'**	
Difference to mean letter count 'm'	Ratio of smaller images	
Difference to mean letter count 'w'	Rel. image position*	
Number of folders in path		
Rel. filename length		
Rel. number of folders in path		

Features in italics are excluded from model training to ensure generality and resilience
*Features marked with asterisks have been introduced by Bender et al. (2016)

In the remainder of the section, we propose a set of features to serve as input to a supervised machine-learning algorithm. Recall that we perform our analysis at the level of an individual image. The features are split into four categories (see Table 10.2). The first two categories, *reference structure* and *HTML image attributes*, subsume aspects that are directly associated with the formatting of the image within the email and its reference URL path. The category *image server* is associated with the servers that host the images. The fourth category covers the *email header*. Features found in prior work (Bender et al., 2016) are marked with an asterisk. While further extension of features is surely possible, we aim to show that a set of resilient features is sufficient to ensure a high level of privacy. We elaborate on the empirical performance of the features in Section 10.7.1.

In the following, we refer to the task of creating features for a detection model as featurization. Featurization is guided by the analysis of the differences in non-tracking and tracking images and domain knowledge regarding the tracking process. We extend the features from Bender et al.

(2016) and select a subset of features for model building based on theoretical considerations of generality and resilience, where resilience describes features and models with stable performance in the event of potential defensive strategies by trackers and changes in tracking infrastructure. It is reasonable to anticipate that companies and tracking providers will adjust their tracking infrastructure to evade anti-tracking efforts; similar to the efforts of spam senders to outsmart spam filters. We expect generality to require features capturing common and inclusive patterns and resilience to require features that cannot be effectively modified by trackers. Two general strategies are applicable to achieve this goal. Based on our understanding of the tracking process, we first exploit the user identifier as an observable and necessary trace of the tracking method and develop features that comprehensively describe its common form as a hash or random letter-number string. The goal is to determine a range of characteristics that are sufficiently general to be prohibitively costly or technically impossible for trackers to avoid. Second, we relate characteristics of single images, which we derive from the data or the related studies on web tracking and ad detection, to other images within the same email. By evaluating each image within the context of the email, potential adjustment strategies by trackers need to consider the infrastructure and conventions used by the content handler. While we engineer and select features based on domain-knowledge and theoretical considerations, future approaches could monitor possible patterns of misclassifications and actual tracker reactions.

10.5.1 Image structure

Image structure features are attributes that are directly associated with an image element and those referring to centrally defined style information from Cascading Style Sheets (CSS). For this category, featurization disregards HTML image attributes occurring in less than 1% of the images in our data set to avoid rare and custom tags and ensure that patterns are detectable and relevant. For tracking images, we expect image attributes to leave display options undefined or make the image harder to detect. For example, a manual inspection of a small set of tracking images suggests the attributes *border* (i.e., the thickness of the border around an image), *style properties* and their respective CSS commands, *vspace* and *hspace*, (i.e., white spaces around images) to have a good discriminatory power (Musciano & Kennedy, 2006).

We further account for the total *number of images* and *relative position* of each image within an email. Our data exploration shows tracking pixels often occur as the first or last image in the email. We suspect that tracking software automatically appends the tracking image to the top or end of an outgoing email to not disturb the email content and furthermore is easier to implement if outsourced tracking services are employed. A second aspect is related to the number of *occurrences of each unique image* within an email. Images used for formatting or branding may be used more than once in one email, but there is no technical nor functional reason to reference the tracking image in an email several times.

A very small image size is often regarded as a typical characteristic of tracking images. There are several ways image size can be specified. The <img> attributes *width* and *height* allow direct specification of the size of the displayed image when the website or email is rendered (Musciano & Kennedy, 2006). Height and width can also be set in the *style* option, sometimes

as a maximum value or in relation to its parent block, or only one dimension can be specified, in which case the image is resized with fixed ratio. It is also possible to not set any size to display the image in its full size. Where no size is explicitly set, we try to extract the image size from the file name, where it is often indicated in the form *image_ 180x120.gif* or similar. Nevertheless, there remain both content and tracking images for which no size information is available, which are classified as "no area specified" (see Figure 10.4). However, there are several theoretical arguments to avoid classification based on image size. First, any image can be tracked independently of size and content. Since there is no technical restriction for tracking images to be of a specific size beyond saving server space, it is likely that tracker will adjust or randomize image size. Second, not all images below a size of 10 square pixels are used for personal tracking. Small or invisible images are also used for the design or formatting of the email content and false classification of these could corrupt display of the email. We consequently exclude all image size features from the models with exception of the ratio of *smaller images within the same email*.

10.5.2 Reference Structure and Content

The majority of the features we propose relate to the referencing link that points to the image (i.e., the URL) with two goals. First, features describe the general structure of the reference to detect patterns that differ from the other image references within the same email, which suggests a third-party tracker. Second, features capture patterns that suggest the existence of a user identifier. Remember that each tracking image reference necessarily contains a unique user ID in the image reference in order to match the image access to a specific email recipient. While the identification of the particular ID of a single user is useful only within the context of the user and the specific sender, there is large potential in features that identify the characteristics of user ID and are resilient to changes by the tracker. In order to capture a range of possible ID structures, we create features that describe the characteristics of the reference path, the content in terms of the reference as a string, and the similarity of each reference to other images in the same email.

The reference structure is captured by a set of features targeting the link folder tree and the characteristics of each of its elements. In addition to the total length and number of elements, we further break up each element in the file path based on punctuation characters. This allows us to collect the characteristics of sub-domains and the referenced files. The observation that the vast majority of tracking images are different from the content images in each mail, which in turn tend to be similar to each other, motivates featurization to capture the *similarity between references in the same email*. We measure link similarity by the Ratcliff/Obershelp text distance between reference URLs (Ratcliff & Metzener, 1988). This text similarity has the property that identical ID tags between references tend to substantially have a high impact on the similarity value due to their relative length. To better capture structural similarity, we additionally quantify the deviation from the majority of references in the same email on several of the features discussed above, including *relative reference length* and *relative path depth.*

A direct approach to flag user identifiers within image references is to blacklist keywords that

indicate tracking functionality. We can identify keywords through text analysis of the references by defining each reference link as a *bag of words* separated by punctuation or special characters and filtered for rare terms. We construct five binary features indicating the existence of tokens that have the highest ratio of occurrence in tracking vs. non-tracking images, such as *uid* or *open,* following the idea is that at least parts of the reference are usually human-readable for convenience. In cases where no random or hashed identifier is used, an @-sign within the URL identifies cases where the email address of the recipient is used as a user identifier directly. While predictive, any specific keywords exist for convenience only and are easily altered or omitted by trackers. Keyword features are consequently excluded from the model features.

A resilient heuristic for ID-like structures is to count the *number of specific special characters* that fulfil a technical role in the tracking infrastructure. In particular, parameters like the user and campaign identifier are passed to tracking scripts through the reference URL. The URL structure required to correctly parse the parameter is defined in public standards (Berners-Lee et al., 2005). The parameters are included behind the file name after a question mark with each key-value pair linked by an equal sign. In contrast to arbitrary keywords, these characters are a necessary component of the tracking infrastructure.

Detection of the structure of identifiers is feasible by counting the occurrences of patterns in the *sequences of upper-/lowercase letters and numbers* and the *distribution of single letters.* These are motivated by the observation that hashes and randomly created image and file names as well as user IDs are expected to contain patterns, e.g. multiple changes in capitalization, and letters that are less common in human-chosen terms. To this end, we create features that capture the difference between how often a letter occurs within a reference to the average number of time the same letter occurs within references of the same email. While these characteristics are within the control of trackers, the design of user IDs which avoid the range of the features requires prohibitive effort. To reduce the set of variables based on letter distribution, we employ preliminary testing using a random forest model on a subset of the data and select the letters *b, f, m,* and *w* as the only predictive letters based on the variable importance score described in Section 10.7.1.

10.5.3 Image Server

External tracking providers regularly host tracking images on their own servers, while content images are likely hosted by the email sender. Even within the same company, we expect images to regularly be provided by different subdomains depending on the process owner. This is supported by our sample, in which more than half of the servers do not host any tracking images. About one-third of the unique domains host tracking images only, while the remaining servers were observed to host both types of images. We capture information on the servers sending the email and hosting the referenced images without restricting the features to specific servers occurring in the collected data. To achieve this, we extract the *ratio of images that share the same domain* and whether the *image host matches the email sender.*

10.5.4 Header Components

An email is composed of an email body and an email header. The email header contains technical information usually not visible to the end user as well as the sender name, address and subject line. An indicator for a *match of the sender name and the image name* aims to capture consistency between the sender and image host. Analysis of the data also shows that a single ID can be used to identify a user or specific message for tracking and to associate unsubscribe requests or email replies with a recipient. In these cases, the respective header fields and the tracking image reference contain an identical ID string. We consequently create features that indicate if parts of each image reference match the content of the header fields *List-unsubscribe, Return-Path,* and *Received-SPF*. These features exploit that one user ID may be used to identify a user in different parts of the infrastructure. While the relevant parts of the sender's infrastructure may lie within the control of the tracker, sufficient changes to the infrastructure will likely be complex and costly.

10.5.5 Server Black-/Whitelisting

Server black- and whitelisting plays a significant role in advertisement and spam detection (Cormack, 2006). In the context of email tracking, the elements of the lists are the image servers that are referenced in the emails. This is an important difference to spam classification, where usually the sender or mail-transfer agent is the object of investigation. Although the data suggest that to block images from servers that have hosted a high ratio of tracking images in the past could be an effective way for identifying tracking images in the data set, the identified servers do not generalize to other companies and potentially not even to one company over a longer period of time. Potential exceptions are third-party tracking services. Since these services take full control of tracking image creation and hosting, their servers show the same pattern, independently of the specific client (see Table 10.3).

To avoid overfitting the classification model to our specific data set, we exclude the identified blacklist and the server locations from model training. Instead, we propose that the images could be filtered according to a black-/whitelist in combination with automated detection or prior to the application of the classification model with the additional benefit of reducing the number of images that need to be classified by the model. The drawback is that these lists are specific, quickly outdated, and require high maintenance effort (Ma et al., 2009b). We use the above list as baseline in the empirical tests below with the caveat that the blacklist could be extended by a comprehensive analysis of the tracking service market in general, which is beyond the scope of this study.

10.6 Methodology

10.6.1 Experimental setup

To verify the effectiveness of the proposed features and the image-classification framework, we empirically test the accuracy of tracking image detection in a real-life environment. This

Table 10.3: Identified tracking service providers and their tracking reference structure

Third-party tracker	**Typical reference structure**
Acxiom Digital	http://open.delivery.net/o?[ID]
Artegic AG	http://[CLIENT].elaine-asp.de/action/view/[ID]/[...]
Conversant (former Dotomi)	http://ads.dotomi.com/cookieredir/[CLIENT[/[...].php?[ID]=1
Doubleclick (Google)	http://ad.doubleclick.net/ad/[...]/[...];ord=[ID];u=[...]?
Mailchimp	http://[CLIENT].[...].list-manage.com/track/open.php?u=[. . .]
Adestra	http://[CLIENT].msgfocus.com/t/[ID].png
MarkMonitor	http://cl.exct.net/open.aspx?[ID]&d=[...]
AppNexus	http://ib.adnxs.com/getuid?http://[...]/[ID]/[...]
Criteo	http://er.prod.verticalresponse.com/[...]/[ID]/pixel.gif
Litmus	https://[CLIENT].emltrk.com/[CLIENT]?d=[MAIL]
Optivo	https://tracking.srv2.de/op/[...]/[ID]-[ID]-[ID].gif
Bigfoot Interactive	http://pix.bfi0.com/t.gif?k=[...]&c=[...]&s=[ID]
Mailermailer	http://m1e.net/c?[ID]
VerticalResponse	http://cts.vresp.com/o.gif?[...]/[ID]/[...]

prepares the development of a fully-functional tracking detection system, in which classification accuracy must be reliable over time and also perform well for senders not included in our sample. We approximate this performance by evaluating classifiers on three dimensions. We report performance on a typical test-set split from the training data, *out-of-sample*, and expand these results with an analysis of two additional test sets. The latter contain newsletters from the same companies sent after the training period, *out-of-time*, and from companies not in the training set, *out-of-universe.* Figure 10.6 summarizes the structure of the training and test setup including the size of the final data sets. The out-of-sample and out-of-universe test sets are drawn randomly from the data collected until October 31, 2015. The out-of-time test sets are emails received in 3-month-periods after the training period. The rest of this section describes the data sets in detail.

The training data consists of HTML emails received within a 5-month period between June 1 and October 31, 2015. It encompasses 215,565 images from 5,478 unique emails. For out-of-sample testing, we randomly select 548 (10%) of these emails and their images to evaluate models trained on the remaining data. The images in the test set are similar to the training data images in that they contain emails from the same time period as the training data and from senders contained in the training data. The sampling process is repeated ten times, respectively. Repeated testing ensures that results are reliable and not due to the random set of emails or companies selected for a single test set. This testing procedure is standard in the machine learning literature and comparable to the approach in Bender et al. (2016).

To construct the out-of-universe test set, we randomly select 30 companies (identified by their sender domain) and assign all of their emails and images to the test set. Since no images from emails of these companies are used to train the model, the corresponding tracking infrastructure and reference structures are entirely unknown to the classifier. Testing on unknown tracking structures allows us to evaluate the performance of the final model on emails sent by different

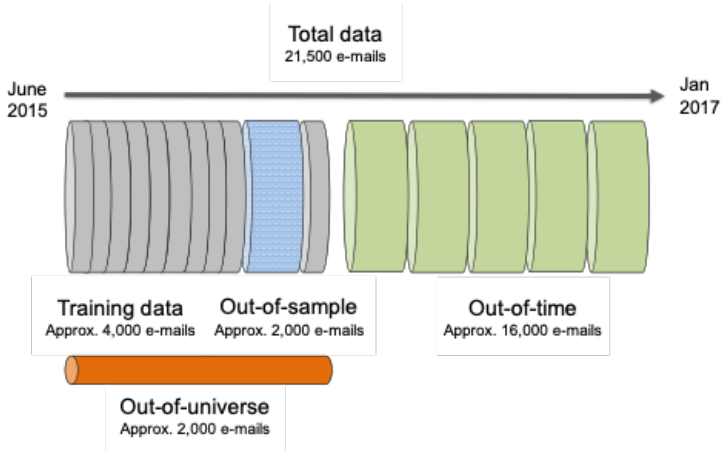


Figure 10.6: Structure and size of the training and three test sets. Note that random sampling of out-of-sample emails and out-of-universe companies is repeated 10 times

companies as an estimate of the performance of the classifiers on unknown senders in a real-world setting. In practice, the ability of the classifier to generalize to tracking infrastructures and senders beyond the 300 companies collected for this research is crucial. The sampling procedure is again repeated 10 times.

In order to capture a potential degradation of classification performance over time, we further define *out-of-time* test sets. The emails received after the training period, i.e. from November 1, 2015 until January 31, 2017, are divided into five sets each of which covers a three-month period. Since the content and structure of the emails and company infrastructure are expected to change over time, the results on the out-of-time sample provide an estimate of the performance of a static classifier after an extended period of time has passed. Since the difficulty of collecting ground-truth data inhibits frequent updating of the model or online learning, robust performance over time is an essential requirement to a reliable blocking approach.

As part of data preparation, we sample a subset of images from the training data for model estimation. The actual distribution of tracking images in the training data, which we use to build binary classifiers, may introduce two forms of bias. First, tracking images make up 8.1% of images in the training data leading to a skewed distribution between the target classes with potentially little variation within non-tracking images in a single email. Unbalanced target classes are known to cause an undesired focus of classification models on the majority class (Verbeke et al., 2012). Second, the numbers of tracking images per email vary from 0 to 57 and thus differ substantially. Hence, tracking images of companies that include a large number of tracking images into their emails may be overrepresented in the data. This may introduce a sampling bias as a classifier may focus on frequent image structures sent by a small number of companies. To overcome these issues, we resample the training data through randomly selecting up to two tracking and content images, respectively, from every email in the training set. For emails containing less than two tracking or content images, respectively, all available images are selected. This approach excludes images from emails with a high number of tracking images and thus addresses the sampling bias. Our resampling also returns an approximately equal amount of tracking and content images for model training. To achieve this, it discards a

sizeable fraction of content images, which suggest that our sampling approach can be considered a form of undersampling (Viaene & Dedene, 2005).

Regarding the application context of tracking prevention, it is important to take into account that the costs associated with different types of errors are uneven. Misclassifying an actual content image as tracking image, and thus filtering the image, may impede readability of the email and negatively influence user experience. On the other hand, misclassifying tracking images, and thus failing to block tracking, impedes user privacy. The proposed selective prevention system is intended to block specific images rather than all images in an email in order to inhibit the user experience as little as possible, while ensuring a maximum level of user privacy. If the cost ratio between false positives and false negatives can be specified, application specific costs can be included into model training, for example through increasing the ratio of target observations in the data via sampling or a reweighting of the model error (Viaene & Dedene, 2005). However, error costs appear an abstract construct in the case of tracking prevention. The misclassification costs depend on the personal risk assessment of an individual user and how she evaluates the relative severity of a privacy breach against the inconvenience associated with manual downloads of blocked images. Given these complications, we argue that a cost-sensitive model estimation is impractical in the focal application context and consider the cost imbalance through post-processing of model predictions described in Section 10.7.2.

10.6.2 Model Selection

We train and test several state-of-the-art machine learning algorithms to identify the binary classifier that is best suited to classify images as "tracking" or "non-tracking" based on the proposed features. Since prior work does not provide information on the performance of classifiers in this application, our selection of methods is based on classifier benchmarks in other domains (Lessmann et al., 2015; Verbeke et al., 2012). All methods take numeric and categorical features as input to identify potentially non-linear patterns and produce a probability estimate of class identity given the feature values for an unknown observation. Each algorithm provides a number of tuning parameters, which describe, for example, the optimization behavior and complexity of the model. Table 10.5 provides a list of candidate models and parameters considered in the study. A comprehensive discussion of the classifiers is beyond the scope of the paper and available in, e.g., (Hastie et al., 2009). We determine the best set of tuning parameters chosen using five-fold cross validation on the training set.

The performance of the state-of-the-art classifiers is compared to three benchmarks. First, we employ a standard logistic regression model. This benchmark allows us to shed light on the trade-off between an interpretable linear model and more complex nonlinear classifiers, which are opaque but supposedly more accurate. Second, we consider the blacklist approach described above as a representative of a manually designed detection rule. This benchmark is to confirm the need for data-driven detection models. Third, we consider a static decision rule based on image size (area below 3 square pixels or not specified) and file format (categories *none*, *php* or *other*). Previous research (Bender et al., 2016) and our exploratory analysis finds these simple features to be highly predictive. It is thus interesting to check the detection performance of

Table 10.5: Classification methods and meta-parameter settings

Learning Algorithms
Artificial Neural Network (Multilayer perceptron) Three-layered architecture of information processing-units referred to as neurons. Each neuron receives an input signal in the form of a weighted sum over the outputs of the preceding layer's neurons. This input is transformed by means of a logistic function to compute the neuron's output, which is passed to the next layer. The neurons of the first layer are simply the covariates of a classification task. The output layer consists of a single neuron, whose output can be interpreted as a class-membership probability. Building a neural-network model involves determining connection weights by minimizing a regularized loss-function over training data. No. of neurons in hidden layer: 3, 5, . . . , 13 Regularization parameter: $2^{[-4,\ -3.5,\ \ldots,\ 0]}$
Random Forest The ensemble consists of fully-grown CART classifiers derived from bootstrap samples of the training data. In contrast to standard CART classifiers that determine splitting rules over all covariates, a subset of covariates is randomly drawn whenever a node is branched, and the optimal split is determined only for these preselected variables. The additional randomization increases diversity among member classifiers. The ensemble prediction follows from average aggregation. No. of member classifiers: 2000 No. of covariates randomly selected for node splitting: 5, 8, 10, 12, 15, 20
Stochastic Gradient Boosting Modification of the AdaBoost algorithm, which incorporates bootstrap sampling and organizes the incremental ensemble construction in a way to optimize the gradient of some differential loss function with respect to the present ensemble composition. We employ tree-based models (CART) as member classifiers. No. of member classifiers: 10, 25, 50, 100, 250, 500 Learning rate: $10^{[-4,\ -3,\ \ldots,\ -1]}$ Max. tree depth: 2, 4, 6, 8

The table depicts only those meta-parameters for which we consider multiple settings. A classification method may offer additional meta-parameters.
We consider all possible combinations of meta-parameter settings for learners such as artificial neural networks that exhibit multiple meta-parameters.

a corresponding classifier and whether it decreases over time. Clearly, the simple classifier, which we refer to as baseline model in the following, is vulnerable to even small adjustments by trackers. The manipulability of image size and displayed file format in particular disqualify this approach as a resilient, long-term solution. For the data employed here, examining the detection performance of the baseline model on the out-of-time data will shed some light on the degree to which an evolution of tracking practices has taken place over the observation period.

10.7 Empirical results

The quality of the detection model depends on its overall performance, generality, and resilience. We measure performance using statistical indicators of predictive accuracy, and generality as model performance under different experimental conditions. In the following, we analyze feature importance to determine the overall number and type of features on which the prediction of a detection model is based and relate these findings to the ease of feature manipulation. We then compare the performance of the models on the different test data sets in terms of their ability to detect tracking images while producing few false alarms. Both characteristics are important

to maximize security and usability for the user, respectively.

10.7.1 Feature importance and resilience

An effective tracking blocker must be able to classify images that vary substantially from the images available for training. Since only a subset of potential senders can be sampled to collect ground-truth data, it is important that features generalize to unobserved senders. Furthermore, the detection engine should be resilient against efforts by trackers to modify their infrastructure to avoid detection. Before discussing the overall performance of the classifiers, we proceed with identifying the salient characteristics of tracking images and evaluate the strength of the proposed new features as determined by the models. Figure 10.7 shows the 15 top-performing features according to normalized feature importance averaged over all classifiers and presents their respective importance values for each classifier.

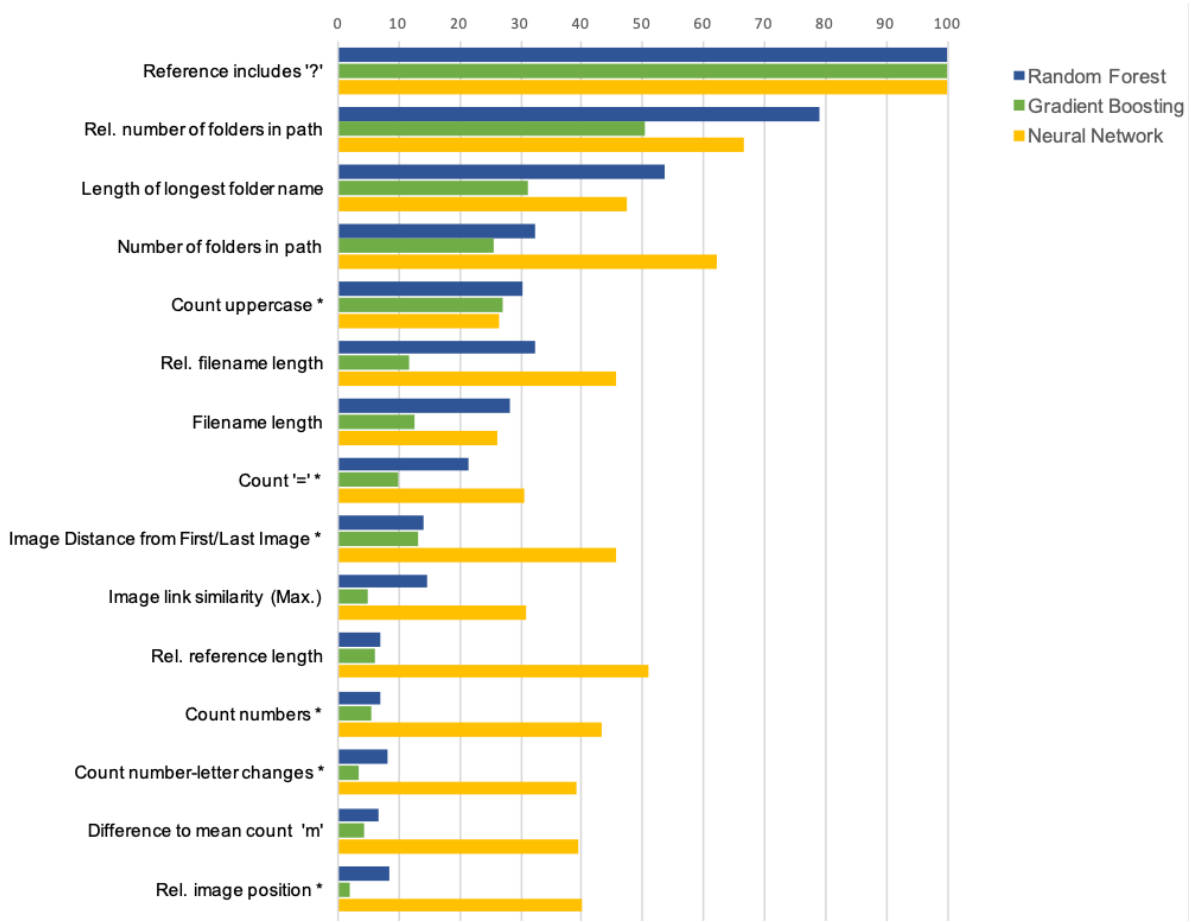


Figure 10.7: The 15 most predictive variables selected according to average feature importance across all classifiers. Features marked with an asterisk have been introduced in Bender et al. (2016)

We use standard algorithm-specific methods to calculate the feature importance scores. For random forest and gradient boosted trees, the score corresponds to the relative improvement in the splitting criterion due to the split; calculated as the square root of the sum of squared relative gain of splits on the feature in a single tree and averaged over all trees in the model (Hastie et al., 2009). For artificial neural network, the hidden-output connection weights of each

hidden neuron are partitioned into components associated with each variable's input neuron using Garson's algorithm (Goh, 1995).

Two main conclusions emerge from Figure 10.7. First, the ranking of variables is similar for all algorithms for the top feature after which there are substantial differences in the ranking between the tree-based models and the artificial neural network. This indicates that there are several highly predictive features within our selection of resilient features. All models rely heavily on the occurrence of a question mark, which indicates that parameters are passed on to a script and the folder structure of the image URL. Beyond the count of uppercase letters, the random forest model seems to consider a larger number of features than the gradient boosting model, e.g. the absolute and relative length of the filename. The neural network distributes importance more evenly and accounts for several defined patterns ignored by the tree-based models, e.g. the count of numbers or case-changes. It also places relatively large weight on email header characteristics, e.g. a match of image and sender domain. In practice, we would expect sparse models to generalize better to unknown data, but models considering a more diverse set of features to be more resilient to changes in tracking patterns.

Second, we observe that the novel features rank high in average importance over all classifiers and make up nine of the fifteen top features. These features are designed for deployability and resilience by capturing patterns that cannot be adjusted without a negative effect on the visibility of the tracking image or its tracking capability. On the side of technical restrictions, the occurrence of a question mark in the reference provides a convenient way to pass parameters to a tracking script and avoiding it would require costly changes to the data collection infrastructure. Even with an alternative solution, the existence of at least one unique identifier is necessary to map the image access to a specific email and email recipient. The existence of these IDs is captured by the (relative) length and number of folders as well as the length of the file name in the top features depicted in Figure 10.7. The large number of recipients requires a certain length and complexity of the ID, which consist of a random number and letter sequence. The randomness of these IDs is captured by the deviation in the number of times a letter is used in each reference to the average occurrence within the email.

On the side of organizationally costly adjustments, changes to the relative number of folders in the URL and all other relative measures regarding the reference structure require flexibility and coordination between different organizational units responsible for the management of content images and tracking images, respectively. For third party trackers, an additional issue is the implementation of changes to the existing server and folder infrastructure adjusted for each client, which requires the restructuring of existing systems. For example, the unification of reference folder length to hide tracking images, which commonly reside in very deep or very shallow folder trees, would require a standardized path structure set by the content management unit that still allows a convenient work environment and does not simultaneously increase the systematic deviation captured by the other relative features, e.g. length of folder name.

Given the high ranking of resilient features designed from technical restrictions and domain-knowledge, we expect the classifiers to be general and resilient with regard to unseen senders

Table 10.6: AUC and average rank classifier performance for each test set (10 sample average)

	Out-of-sample			Out-of-universe			Out-of-universe & -time		
	AUC	**Rank**		**AUC**	**Rank**	**AUC**	**Rank**		
Blacklist	0.596	6.00	*(0.00)*	0.673	6.00	*(0.00)*	0.637	6.00	*(0.00)*
Baseline	0.969	5.00	*(0.00)*	0.953	5.00	*(0.00)*	0.938	4.80	*(0.00)*
Logit	0.998	4.00	*(0.04)*	0.982	4.00	*(0.03)*	0.972	3.95	*(0.03)*
NN	1.000	2.10	(1.00)	0.994	2.40	(0.95)	0.981	2.55	(0.68)
RF	1.000	1.95	–	0.997	1.80	–	0.994	1.75	–
GBT	1.000	1.95	(1.00)	0.996	1.80	(1.00)	0.993	1.95	(0.81)
Friedman χ^2_5		49.57	*(0.00)*		48.43	*(0.00)*		44.65	*(0.00)*

Values in brackets give the adjusted p-value corresponding to a pairwise comparison of the row classifier to the best classifier (random forest). Italic face indicates significance at the five percent level. The last row shows the χ^2 and p-values of a Friedman test to verify that at least two classifiers perform significantly different.
NN: Neural Network, RF: Random Forest, GBT: Gradient Boosting

and changes over time, and against expected deliberate changes in the tracking infrastructure as outlined above. The following section evaluates the former two claims empirically.

10.7.2 Model performance

We evaluate classifier performance based on the area-under-the-ROC-curve (AUC), which captures a classifier's ability to discriminate between tracking and non-tracking images. We also use sensitivity and specificity statistics based on the optimal probability threshold to evaluate the tracking detection accuracy of a classifier vis-à-vis its ability to not block content images.

The AUC allows us to summarize the performance of each classifier in a single metric aggregated over all potential thresholds and test the differences in performance statistically. The AUC for each classifier and test set, averaged over ten repetitions of random sampling, is given in Table 10.6. Note that the AUC is bounded between 0 and 1 (perfect discrimination), where a value of 0.5 corresponds to a random classifier. We also report the average ranks as the basis of a statistical analysis of model performance comparing the classifiers to the best performing classifier (Demšar, 2006). The last row of Table 10.6 depicts the test statistic and p-value of a Friedman test of the null-hypothesis that all classifier ranks are equal. Given that we can reject the null-hypothesis for all performance measures ($p < .00$), we proceed with pairwise comparisons of a classifier to the control classifier using the Rom procedure for p-value adjustment (García et al., 2010). Table 10.6 depicts the p-values corresponding to the pairwise comparisons in brackets. Italic face indicates that we can reject the null-hypothesis of a classifier performing equal to the best classifier (i.e., $p < .05$).

For all test sets, the random forest model performs best and thus serves as control model for statistical testing. We observe the benchmark models to perform significantly worse than the random forest classifier at the 5% level but are unable to establish a significant difference in performance between the random forest and the gradient boosting or neural network classifier. The results for the out-of-sample test set are comparable to previous studies and support the

view that machine-learning classifiers are highly effective in identifying tracking elements (Bender et al., 2016). All non-linear classifiers achieve close to perfect performance and perform significantly better than the baseline model, which classifies images based on image size and file format. The blacklist model provides some discriminatory power. However, it performs significantly worse than the best alternative classifier.

Out-of-sample results represent the performance of a detection engine under ideal conditions. In practice, we cannot expect emails to originate from the same senders as in the training data. Additionally, the challenges in collecting labelled training data restrict model training to a relatively small number of different senders and impede regular re-training or updating of classifiers. We therefore evaluate the classifiers on out-of-universe test cases, which include only images from companies on which the model was not trained, and out-of-universe-and-time test cases, which include images from companies on which the model was not trained received after the end of the training period. As expected, we observe a decrease in AUC for all classifiers when applied to the more challenging test sets. This decrease is lowest for the random forest and gradient boosting models at 0.006 and 0.007 AUC points and highest for the logit model with a difference of 0.026 AUC points, suggesting that the tree-based ensemble models generalize well. Decreasing performance of the baseline model is surprising given its simple decision rule set and hints at a change in tracking practices in the out-of-universe/-and-time test sets. We attribute the fact that AUC increases for the blacklist model on more challenging test sets to sampling variance. The performance of the blacklist model is high whenever companies that use trackers from the blacklist are sampled for a random test set, and low otherwise.

The excellent discriminatory performance of classifiers, even on the out-of-universe-and-time test set, facilitates two conclusions. First, the proposed features are sufficient to allow near perfect classification of tracking images within newsletter emails. This is important empirical validation that it is possible to identify tracking images without relying on image characteristics that are controlled by trackers. AUC values close to unity suggest that a tracking detection system built on resilient features can provide effective protection in the long run. Second, the proposed classification models generalize to newsletter emails received after the training period and from unknown companies. Generalizability is crucial due to the high number of potential senders and the discussed difficulties in data collection, which impede frequent updating of the detection model.

Having established the predictive performance of the detection models, we examine the binary decision between loading and blocking an image in practice. This requires us to post-process the probabilistic predictions emerging from classification models. We obtain a crisp classification of images into tracking and non-tracking images through comparing probabilistic classifier predictions to a threshold. We then assess the accuracy of discrete class predictions in terms of the sensitivity and specificity of a classifier, i.e. the percentage of tracking and non-tracking images that are correctly classified, respectively. The definition of a threshold also offers an opportunity to account for uneven misclassification costs without having to specify actual cost values. We tune the probability threshold (for each classifier individually) on the training data set. Similar

Table 10.7: Sensitivity and specificity of detection models across ten random test sets

	Out-of-universe & -time										
	1	2	3	4	5	6	7	8	9	10	**Mean**
					Sensitivity						
Blacklist	0.03	0.19	0.35	0.23	0.47	0.28	0.15	0.29	0.41	0.51	29%
Baseline	1.00	0.81	0.96	0.96	0.89	0.96	0.96	0.95	0.91	0.98	94%
Logit	1.00	0.98	0.99	1.00	0.93	0.99	0.90	1.00	0.99	1.00	98%
Neural Network	1.00	0.98	0.78	0.82	0.98	0.73	0.30	0.97	0.79	0.97	83%
Random Forest	1.00	0.99	0.82	0.81	0.99	0.86	0.77	0.96	1.00	0.99	92%
Gradient Boosting	1.00	1.00	0.86	0.85	0.84	0.88	0.80	0.96	1.00	0.99	92%
					Specificity						
Blacklist	0.99	1.00	0.98	0.98	0.98	0.98	0.95	1.00	0.96	1.00	98%
Baseline	0.94	0.94	0.96	0.92	0.95	0.90	0.95	0.93	0.93	0.98	94%
Logit	0.54	0.93	0.78	0.59	0.80	0.74	0.89	0.76	0.56	0.59	72%
Neural Network	0.96	0.98	1.00	0.99	0.98	0.98	1.00	0.99	0.94	0.99	98%
Random Forest	0.90	0.98	1.00	0.99	0.97	0.99	1.00	0.99	0.95	0.98	98%
Gradient Boosting	0.96	0.91	1.00	0.99	0.98	0.99	1.00	0.99	0.85	0.99	97%

to applications in spam detection (Bergholz et al., 2008), medicine (Oztekin et al., 2017) and fraud detection (Van Vlasselaer et al., 2017; Viaene et al., 2007), the goal is to achieve a high detection rate with the lowest possible rate of false alarms. For the empirical evaluation, we define the probability threshold to be the value that maximizes the specificity of a classifier at a fixed sensitivity of at least 99.99% on the training data. We acknowledge the choice of 99.99% to be subjective. It is based on the believe that many users might have a strong preference for privacy and consider the misclassification of a tracking image to be the much more "costly" error compared to misclassifying a content image. Having fixed the sensitivity of each model on the training data, we compare the models on the most challenging out-of-universe-and-time scenario by comparing the sensitivity and specificity over ten random samples. In practice, sensitivity and specificity correspond to the ratio of detected tracking images and one minus the ratio of (erroneously) blocked non-tracking images, respectively. Results are presented in Table 10.7.

For all classifiers, we observe sensitivity to differ from our target value of 99.99%. Recall that this is the target value which we use to determine the classification threshold on the training data. Table 10.7 demonstrates that applying this threshold to unknown data decreases sensitivity (i.e., the accuracy of tracking image detection). Considering the tradeoff between high sensitivity and a low false alarm rate, Table 10.7 reveals that the random forest classifier has a higher tendency to sacrifice sensitivity for higher specificity compared to the logit model. We attribute the sharper decrease in sensitivity for random forest to the fact that random forest achieves almost perfect discrimination on the training data (Table 10.6), which leads to a higher, less strict classification threshold after optimization. Overall, the results suggest that the excellent discriminatory power observed in terms of AUC (Table 10.6) translates well to the actual decision problem under the proposed cutoff optimization scheme. When a user decides to allow loading external images for an email, the logistic regression or random forest classifiers robustly detect 98% and 92% of tracking images under the proposed system. At this level of performance, the detection models ensure a high level of user privacy under the most challenging

conditions of an out-of-universe-and-time test. The negative effect on user experience is the false flagging of 28% and 2% of non-tracking images as tracking images, respectively. In the case of the random forest, we argue that the privacy gain outweighs the negative effect for users with even minimal preference for privacy. We judge the logistic regression under the proposed cutoff to be an alternative for users with a strong preference for privacy at the cost of a notable impact on user experience.

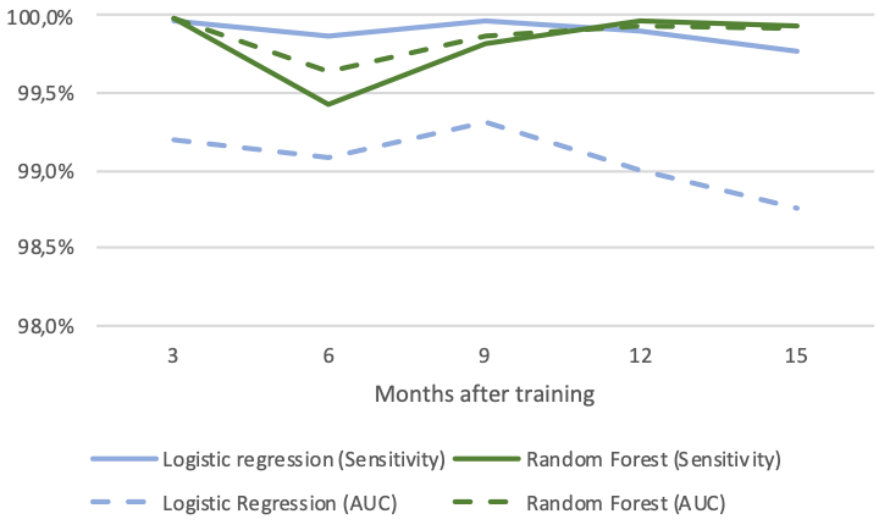


Figure 10.8: Sensitivity after training period over five 3-month windows

We further compare the dynamics of AUC and sensitivity of the selected models over time and determine a suitable interval for retraining each model. We conduct this analysis over five 3-month windows starting at the end date of the training data on out-of-sample test data to ensure sufficient sample size in each window. Figure 10.8 shows no trend in performance for the random forest and only marginal decrease in performance for logistic regression starting after nine months, with all changes within a 1%-interval of starting performance. We attribute the slump in performance for the 3-6-month window to particularities in the email schedule for the subscribed newsletter. Taken at face value, the results suggest that a detection engine preserves performance for a period of at least nine months after training, at which point the logistic regression could be re-trained on more recent data to avoid deteriorating performance. No long-term time effect is observed for the random forest within the observed period. These results are encouraging for practical applications where data collection and retraining are challenging. Furthermore, higher robustness of the random forest model supports the view that this model might be preferable to the logistic regression despite its lower sensitivity.

10.8 Conclusion

Email tracking can be used to gather identifiable and sensitive information on recipients without their consent or control, thus raising several security and privacy concerns. We describe the extent to which data can be collected that contains information on email reading behavior, system information, and location. In contrast to common web tracking, these data can be matched to an email address and, by extension, to the person behind the email account.

Empirical analysis of over 30,000 emails from the 100 largest companies in Germany, Great Britain, and the United States, respectively, show that email tracking is widely applied. About 50% of all newsletter emails and close to 100% of emails in consumer-oriented industries include at least one tracking image. We identify the lack of a general, reliable and sufficient protection system against email tracking in previous literature and the software market, and propose a selective-prevention solution on the image level that is most suitable to balance privacy and usability.

We use the collected data to build a detection engine for the identification of tracking images based on machine learning. To achieve this, we outline a general methodology to infer resilient features from the technical characteristics of the tracking process. We follow this approach to design a comprehensive set of features that ensure applicability and resilience against tracker counter-strategies in a real-world setting. We test three state-of-the-art machine-learning classifiers and benchmark expected performance against heuristics proposed in previous research in a realistic application setting. In particular, we take into account long-term changes of tracking structures and classification of emails from unknown senders through repeated random sampling of three test data sets. We find a random forest classifier to provide the best overall classification performance at a detection rate of 92% and misclassification rate of non-tracking images at 2% for newsletters received from unknown senders after the training period.

Some caveats apply to the results gathered in this study, indicating directions for future research. The data used in this study contains commercial email newsletters, which exhibit important advantages for this research setting. Typical mail use involves additional mail categories, including private messages and the large category of spam and phishing emails. Further studies are required to check if our results generalize to tracking mechanisms in different types of email contexts. A fundamental threat to tracking image detection systems comes from the risk that actual content images could be employed for tracking. The proposed model could detect such images. However, their removal or blocking has a direct impact on the informational content of the email and thus conflicts with the interest of the user. Content-image tracking could be addressed using a server-side proxy solution. The server could cache all images with high tracking probability. Subsequent access to these (content and tracking) images from email recipients can then reference the server. Trackers would observe image downloads but only from the server so that the privacy of individual users is not compromised. The role of the proposed detection engine in a server-side solution would be to improve efficiency. Through selecting likely tracking images the server does not have to cache all images in all incoming emails. Furthermore, the problem of using content images for tracking is mitigated by the fact that tracking applications are often provided by specialized third-party services, for which the implementation of tracking mechanisms to content images within an email would require far more effort than attaching content-less tracking images. With such separation of content-management and tracking, an integrated solution will be costly for companies to realize. Further research on user behavior will prove useful to determine if users are willing to manually allow loading tracked content images.

With an extension of the data collection period, an analysis of changes to the features employed by the models and monitoring of model performance over time may provide insights into developments in tracking infrastructure and active countermeasures. Taking the long-term perspective, we have outlined the strategies that are available to trackers in order to actively hide tracking images from simple detection heuristics. Based on the available data and observation period, we come to the conclusion that the proposed detection system performs effectively and stable on the basis of the proposed resilient features. Consistently high tracking image detection rates on out-of-time and out-of-universe data suggest that the distribution of feature values or tracking practices has not changed during the observation period. Such change may however occur in the future. Therefore, future research to replicate our results and to perform a longitudinal analysis of feature distributions to collect evidence for a potential distributional shift seems highly relevant.

Bibliography

Alsaid, A., & Martin, D. (2003). Detecting Web Bugs with Bugnosis: Privacy Advocacy through Education (R. Dingledine & P. Syverson, Eds.). In R. Dingledine & P. Syverson (Eds.), *Privacy Enhancing Technologies*, Berlin, Heidelberg, Springer. https://doi.org/10.1007/3-540-36467-6_2

Barrett, B. (2015). A Clever Way to Tell Which of Your Emails Are Being Tracked. *Wired*.

Bender, B., Fabian, B., Lessmann, S., & Haupt, J. (2016). E-Mail Tracking: Status Quo and Novel Countermeasures, In *Proceedings of the 37th International Conference on Information Systems (ICIS)*, AIS.

Bergholz, A., Chang, J. H., Paass, G., Reichartz, F., & Strobel, S. (2008). Improved Phishing Detection using Model-Based Features, In *Proceedings of the Fifth Conference on Email and Anti-Spam*.

Berners-Lee, T., Fielding, R., & Masinter, L. (2005). *Uniform Resource Identifier (URI): Generic Syntax* (RFC No. 3986). RFC Editor.

Blum, A., Wardman, B., Solorio, T., & Warner, G. (2010). Lexical Feature Based Phishing URL Detection Using Online Learning, In *Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security*, ACM.

Bonfrer, A., & Drèze, X. (2009). Real-time evaluation of e-mail campaign performance. *Marketing Science*, *28*(2), 251–263. https://doi.org/10.1287/mksc.1080.0393

Bouguettaya, A., & Eltoweissy, M. (2003). Privacy on the web: Facts, challenges, and solutions. *IEEE Security & Privacy Magazine*, *1*(6), 40–49. https://doi.org/10.1109/MSECP.2003.1253567

Cormack, G. V. (2006). Email spam filtering: A systematic review. *Foundations and Trends® in Information Retrieval*, *1*(4), 335–455.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

Dobias, J. (2011). Privacy Effects of Web Bugs Amplified by Web 2.0, In *Privacy and Identity Management for Life*, Berlin, Heidelberg, Springer. https://doi.org/10.1007/978-3-642-20769-3_20

Englehardt, S., Han, J., & Narayanan, A. (2018). I Never Signed Up For This! Privacy Implications of Email Tracking, In *Proceedings on Privacy Enhancing Technologies*.

Evans, M., & Furnell, S. (2003). A model for monitoring and migrating web resources. *Campus-Wide Information Systems*, *20*(2), 67–74. https://doi.org/10.1108/10650740310467763

Evers, J. (2006). How HP Bugged E-Mail: Commercial Online Service Was Used to Track E-mail Sent To a Reporter in Hewlett-Packard's Leak Probe, Investigator Testifies. *CNET*.

Fette, I., Sadeh, N., & Tomasic, A. (2007). Learning to Detect Phishing Emails, In *Proceedings of the 16th International Conference on World Wide Web*, ACM.

Financial Times. (2017). Equities. *Financial Times*.

Fonseca, F., Pinto, R., & Meira, W. (2005). Increasing User's Privacy Control through Flexible Web Bug Detection, In *Third Latin American Web Congress*, Buenos Aires, Argentina, IEEE. https://doi.org/10.1109/LAWEB.2005.19

García, S., Fernández, A., Luengo, J., & Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, *180*(10), 2044–2064.

Garera, S., Provos, N., Chew, M., & Rubin, A. D. (2007). A Framework for Detection and Measurement of Phishing Attacks, In *Proceedings of the 2007 ACM workshop on Recurring Malcode*, ACM.

Goh, A. T. C. (1995). Back-propagation neural networks for modeling complex systems. *Artificial Intelligence in Engineering*, *9*(3), 143–151. https://doi.org/10.1016/0954-1810(94)00011-S

Harding, W. T., Reed, A. J., & Gray, R. L. (2001). Cookies and web bugs: What they are and how they work together. *Information Systems Management*, *18*(3), 17–24.

Hasouneh, A. B. I., & Alqeed, M. A. (2010). Measuring the effectiveness of e-mail direct marketing in building customer relationship. *International Journal of Marketing Studies*, *2*(1), 48–64. https://doi.org/10.5539/ijms.v2n1p48

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The Elements of Statistical Learning* (Second). New York, Springer.

Hlatky, P. (2013). The Yesware Follow Up.

Hodgekiss, R. (2010). How Do I Create a Printer-Friendly Email Newsletter?

Javed, A. (2013). POSTER: A Footprint of Third-Party Tracking on Mobile Web, In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, Berlin, Germany, ACM. https://doi.org/10.1145/2508859.2512521

Jensen, C., Sarkar, C., Jensen, C., & Potts, C. (2007). Tracking Website Data-collection and Privacy Practices with the iWatch Web Crawler, In *Proceedings of the 3rd Symposium on Usable Privacy and Security*, Pittsburgh, PA, USA, ACM Press. https://doi.org/10.1145/1280680.1280686

Kan, M.-Y., & Thi, H. O. N. (2005). Fast Webpage Classification using URL Features, In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, ACM.

Kushmerick, N. (1999). Learning to Remove Internet Advertisements, In *Proceedings of the 3rd Annual Conference on Autonomous Agents*, ACM.

Leon, P., Ur, B., Shay, R., Wang, Y., Balebako, R., & Cranor, L. (2012). Why Johnny Can't Opt Out: A Usability Evaluation of Tools to Limit Online Behavioral Advertising, In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Austin, Texas, USA, ACM Press. https://doi.org/10.1145/2207676.2207759

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: Online Appendix. *European Journal of Operational Research*, *247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Li, M., Li, Z., Li, D., & Wang, B. (2010). *Classification of images as advertisement images or non-advertisement images* (EP2203870A2).

Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2009a). Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs, In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM.

Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2009b). Identifying suspicious URLs: An application of large-scale online learning, In *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM.

Martin, D., Wu, H., & Alsaid, A. (2003). Hidden surveillance by web sites: Web bugs in contemporary use. *Communications of the ACM*, *46*(12), 258. https://doi.org/10.1145/953460.953509

Moscato, D. R., Altschuller, S., & Moscato, E. D. (2013). Privacy policies on global banks' websites: Does culture matter? *Communications of the IIMA*, *13*(4), 91–109.

Murphy, K. (2014). Ways to Avoid Email Tracking. *New York Times*.

Musciano, C., & Kennedy, B. (2006). *HTML & XHTML: The Definitive Guide*. O'Reilly Media, Inc.

Nikiforakis, N., Kapravelos, A., Joosen, W., Kruegel, C., Piessens, F., & Vigna, G. (2013). Cookieless Monster: Exploring the Ecosystem of Web-Based Device Fingerprinting, In *Proceedings of the 2013 IEEE Symposium on Security and Privacy*, IEEE. https://doi.org/10.1109/SP.2013.43

Oztekin, A., Al-Ebbini, L., Sevkli, Z., & Delen, D. (2017). A decision analytic approach to predicting quality of life for lung transplant recipients: A hybrid genetic algorithms-based methodology. *European Journal of Operational Research*. https://doi.org/10.1016/j.ejor.2017.09.034

Poese, I., Uhlig, S., Kaafar, M. A., Donnet, B., & Gueye, B. (2011). IP geolocation databases: Unreliable? *ACM SIGCOMM Computer Communication Review*, *41*(2), 53. https://doi.org/10.1145/1971162.1971171

Premkumar, G., & Roberts, M. (1999). Adoption of new information technologies in rural small businesses. *Omega*, *27*(4), 467–484.

Ratcliff, J. W., & Metzener, D. E. (1988). Pattern Matching: The Gestalt Approach. *Dr. Dobb's Journal*, *13*(7), 46.

Shih, L. K., & Karger, D. R. (2004). Using URLs and Table Layout for Web Classification Tasks, In *Proceedings of the 13th International Conference on World Wide Web*, ACM.

Technology Analysis Branch. (2013). *What an IP Address Can Reveal About You* (tech. rep.). Office of the Privacy Commissioner of Canada.

The Direct Marketing Association. (2015). *National Client Email Report 2015* (tech. rep.).

Vaas, L., & Stockley, M. (2014). How Emails Can Be Used to Track Your Location and How to Stop It.

Van Vlasselaer, V., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2017). GOTCHA! Network-based fraud detection for social security fraud. *Management Science*, *63*(9), 3090–3110. https://doi.org/10.1287/mnsc.2016.2489

Vaynblat, D., Makagon, K., & Tsemekhman, K. (2009). *System and Method for Automatically Delivering Relevant Internet Content* (US20100082808A1).

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, *218*(1), 211–229. https://doi.org/10.1016/j.ejor.2011.09.031

Viaene, S., Ayuso, M., Guillen, M., Van Gheel, D., & Dedene, G. (2007). Strategies for detecting fraudulent claims in the automobile insurance industry. *European Journal of Operational Research*, *176*(1), 565–583. https://doi.org/10.1016/j.ejor.2005.08.005

Viaene, S., & Dedene, G. (2005). Cost-sensitive learning and decision making revisited. *European Journal of Operational Research*, *166*(1), 212–220.

Whittaker, C., Ryner, B., & Nazif, M. (2010). Large-Scale Automatic Classification of Phishing Pages, In *The Network and Distributed System Security Symposium NDSS'10*.

Yang, Y. ((2010). Web user behavioral profiling for user identification. *Decision Support Systems*, *49*(3), 261–271. https://doi.org/10.1016/j.dss.2010.03.001

Chapter 11

Enterprise-Grade Protection Against E-Mail Tracking

Publication
Fabian, B., Bender, B., Hesseldieck, B., Haupt, J., & Lessmann, S. Enterprise-Grade Protection Against E-Mail Tracking. Under review at Information Systems.

Abstract
This article presents a novel protection framework against e-mail tracking that closes an important gap in the field of enterprise security and privacy-enhancing technologies. We conceptualize, implement and evaluate a professional anti-tracking mail server that is capable of identifying tracking images in e-mails via machine learning with very high accuracy, and can *selectively* replace them with arbitrary images containing warning messages for the recipient. Our mail protection framework is developed as enterprise-grade software using the design science research paradigm. It is flexibly extensible, highly scalable, and ready to be applied in actual production conditions. Experimental evaluation shows that these goals are achieved through solid software design, adoption of recent technologies and the creation of novel flexible software components.

11.1 Introduction

Online marketing has grown to a 70 billion dollar industry worldwide annually (Farahat & Shanahan, 2013). It comprises all media that make use of the Internet, as for example Web, social media and e-mail, and aims at promoting products or brands to a certain target group. Despite its impressive growth, online marketing encounters many challenges such as how to determine the efficiency or the loss of sales caused by the incorrect use of advertising budget. Thus, it seems important for firms to measure, maximize, and benchmark the effectiveness of advertising compared to media expenditures, for optimizing efficiency (Farahat & Shanahan, 2013). To achieve this, marketers made use of the technologies behind the media, which resulted in one of the competitive advantages of online marketing: tracking.

Tracking methods, such as web and e-mail tracking, are popular marketing tools due to the increasing importance of accurate customer data for business success (Ermakova et al., 2017; Goldfarb & Tucker, 2012) Analysis of customer data is used to personalize offerings and marketing layouts for an optimized market position as well as an advantage in product pricing (Ridley-Siegert, 2016). Data derived from tracking provides valuable information regarding a

person's interests and reception behavior. Modern e-mail tracking methods allow the sender to determine how often an e-mail was opened, the device used to read the e-mail, which links were clicked, and the location and time when the recipient opened an e-mail (Englehardt et al., 2018; Fabian et al., 2015).

The breach of privacy caused by tracking is severe, not just because of the fine data granularity but also due to the fact that the data were gathered without permission, request, and often knowledge, from the customer. Tracking data are so valuable that companies specialize in their gathering and use selling of aggregated-data as a business model (Li et al., 2015) or offering e-mail tracking as a service. Furthermore, hackers and criminals use e-mail tracking to determine if an e-mail account is active and whether the owner opens attachments, which paves the way for system intrusions (Vaas & Stockley, 2014). E-mail tracking can thus be seen as a serious privacy and security threat for end-users and companies.

Prior research on e-mail tracking examined the principles (Englehardt et al., 2018; Fabian et al., 2015) as well as the usage across regions (Bender et al., 2016) and the actual employment of data gathered through e-mail tracking (Bender et al., 2018). Haupt et al. (2018) examined the potential of machine-learning to identify tracking elements in e-mail communication. Their results indicate that accurate detection is possible, which may be seen as a first step towards tracking prevention and privacy protection. Achieving the latter, however, requires the implementation of the approach considered by Haupt et al. (2018) in their laboratory experiment that considers real-world requirements related to the efficiency and scalability of a detection engine, amongst others. To the best of our knowledge, reliable software for tracking prevention as well as empirical insights how detection systems behave under realistic deployment conditions is underrepresented in existing literature. This study aims to close this research gap in the field of privacy-enhancing technologies, by conceptualizing and implementing a novel protection framework against e-mail tracking. We realize a professional anti-tracking mail server that is capable of identifying tracking images in e-mails via machine learning and selectively replace them to mitigate tracking. Our mail protection framework is developed as enterprise-grade software, flexibly extensible, highly scalable, and ready to be applied in actual production conditions. The experimental evaluation section shows that this is achieved through corresponding choices regarding technologies and the creation of a solid software design.

Our study follows the paradigm of design science research (Peffers et al., 2007). The structure of this article is aligned with the major steps of this established process, starting with the problem identification and motivation in this introductory section. In the next section, the required research background is presented in order to prepare the objectives and requirements in the third section. Then, the design and development process of our software artifact is elaborated, followed by a demonstration and thorough experimental evaluation. Finally, major contributions, limitations and future work are discussed.

11.2 Related Work

Publications on the general issue of web tracking are numerous, ranging from its usage (Ermakova et al., 2018; Ermakova et al., 2017; Javed, 2013; Jensen et al., 2007; Mittal, 2010; Parra-Arnau, 2017) to its detection (Alsaid & Martin, 2003; Fonseca et al., 2005; D. Martin et al., 2003; Yamada et al., 2011) and prevention (Bujlow et al., 2017; Fonseca et al., 2005; Leon et al., 2012; Sanchez-Rola et al., 2017). Various publications also hint at the opportunity to apply the tracking mechanics used in the web for HTML-based e-mail (Bouguettaya & Eltoweissy, 2003; Bujlow et al., 2017; Harding et al., 2001; Li et al., 2015; D. Martin et al., 2003). However, none of these studies explore this possibility further.

E-mail tracking itself is mainly investigated and utilized in the context of at least four research fields: marketing, malicious e-mails, spam, and privacy. The first research stream explores the effects of *e-mail marketing* on the recipients and its optimization (Bilos et al., 2016) (Hartemo, 2016; Luo et al., 2015; Zhang et al., 2017), while utilizing data gathered through e-mail tracking. Publications from Bonfrer and Drèze (2009) as well as from Hasouneh and Alqeed (2010) investigate the tracking technology and process from a marketing viewpoint and emphasize the importance of tracking newsletters and any other e-mail marketing communication. The Direct Marketing Association (DMA) releases annually a thirty-sided research report about e-mail tracking and its impact on online marketing (Ridley-Siegert, 2016).

The second research avenue is the field of *malicious e-mail* content or attachments, where e-mail tracking is used to analyze the mechanics and velocity of spreading viruses or malicious programs (Bhattacharyya et al., 2002; Stolfo et al., 2003). Third, the issue of *spam* is necessarily an important issue in e-mail research, but also e-mail tracking has connections to it. Mechanics of e-mail tracking are used to identify the origin of spam mail in order to add the sender to blacklists (Grimes et al., 2007; Hameed et al., 2013; Herzberg, 2009). The fourth field this article draws upon is *privacy*. Protecting privacy against tracking that aims to expose the end-user's personal information to marketers is the main focus of this research direction. Various environments are investigated, such as the company level, the browser, and e-commerce (Englehardt et al., 2018; Fabian et al., 2015; Gu et al., 2017; Sammons & Cross, 2017; Tsalis et al., 2016). While prior studies considered only the technical feasibility of e-mail tracking, Bender et al. (2018) showed the actual use of customer-behavior data.

Although the issue of e-mail tracking is analyzed in literature, no effort has been devoted to the creation of a software solution for end-users, nor are prevention methods explored in depth, with the exception of Bonfrer and Drèze (2009) and Englehardt et al. (2018). The former focused on investigating the current status of e-mail tracking and the design space for effective countermeasures and also provided initial drafts and evaluation of a machine-learning based detection model. Englehardt et al. (2018) empirically surveyed the current landscape of mail tracking with a particular emphasis on third-party trackers. Furthermore, they assessed the incompleteness of existing defense solutions, and briefly outlined a novel anti-tracking strategy based on filter lists. Additionally, Haupt et al. (2018) benchmarked various machine-learning approaches and thereby guided the design of effective and reliable detection mechanisms.

These results provide the motivation for the countermeasures and detection model utilized by this study's software artifact. Sharing the goal of preserving e-mail end-users privacy, this article aims to implement a comprehensive and enterprise-grade countermeasure solution.

11.3 Solution Objectives

The paper aims to provide results regarding the research question: How can reliable and efficient anti-tracking mechanisms be realized in enterprise-level environments? As a first step to provide a sufficient protection mechanism for e-mail tracking in an enterprise context, requirements should be gathered and objectives need to be set.

Planning any software project starts with the decisions on what the software should accomplish (Sommerville & Sawyer, 1997). The solution proposed here needs to satisfy the following functional requirements. First and foremost, following the preventive selection approach, the software should provide a reliable detection mechanism that identifies tracking images in HTML-based mails and replaces them with an arbitrary non-tracking image, e.g., a fixed warning sign to inform the user of the tracking attempt. This involves the development of a detection engine, which after a training period can automatically classify images into tracking and non-tracking. Training the engine can be conducted offline before deployment, or online via updates during server operation. Second, fully-fledged mail server functionality and a link replacement engine are required.

Concerning technical requirements, framework modularity is to be named for easy maintainability and extensibility. For example, enabling developers to plug-in new machine learning models and algorithms into the framework. Another requirement is stability: providing a framework that is built on modern but well-tested technologies is crucial. Not least is scalability: Large companies or e-mail application provider quickly deal with thousands of mails in a short timeframe. This demands an elastic solution, supporting small-scale use cases while being able to scale up to an enterprise level. As hard limits, we decided that the implementation of the framework should not require more than 500 megabyte (MB) storage space and less than 250 megabyte of memory (both without the detection engine) since these resources can be assumed to be available in most enterprise infrastructures or included using cloud services, where such configurations are usually offered for free. Further, a reasonable buffer should be planned for both metrics, in order to be prepared for extensions and peak traffic situations.

For practical enterprise applications, performance is important. On the one hand, this is not extremely crucial in the field of e-mails since it usually does not need to work in real-time. On the other hand, poor performance leads to limited scalability and high use of resources. An additional processing of e-mails by the proposed tracking prevention system will inevitably cause some delay in e-mail delivery. We suggest a maximum threshold of five minutes for practicability reasons. The system should keep the delay in the range of seconds but under no circumstances exceed the five-minute threshold. Regarding scalability, the framework should be able to cope with at least 20 concurrent connections. Achieving such a degree of scalability qualifies the framework for enterprise applications, while the explicitly low resource requirements allow the

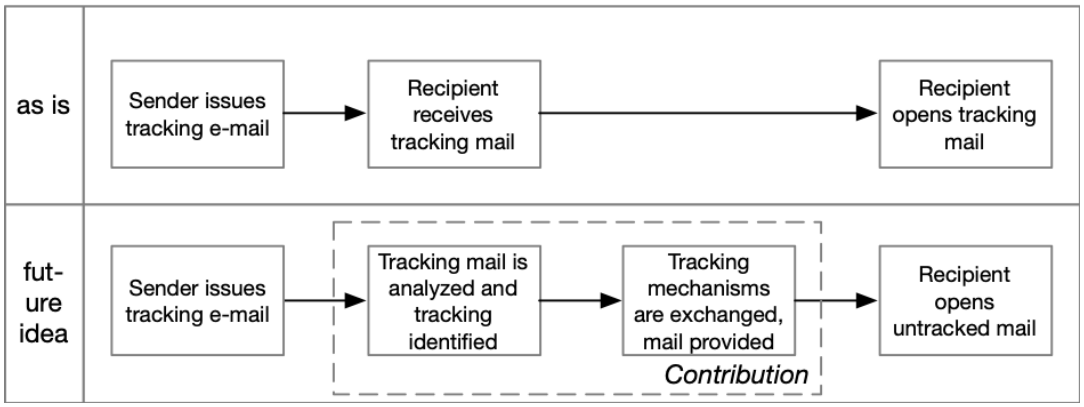


Figure 11.1: Comparison *as is* and desired tracking approach

system to be run on free popular Platform-as-a-Service providers.

Figure 11.1 shows the unprotected *as-is* and future process scenarios for tracking prevention. Furthermore, it subsumes the major components (detection and exchange) and their related modularity for the solution.

11.4 Design & Development

Designing a solution as shown in the lower panel of Figure 11.1 requires taking several decisions. First, the tracking approach needs to be defined. Afterwards high-level realization decisions need to be taken in a second step. Finally, in a third realization step, necessary decisions regarding architecture and implementation need to be taken.

Potential tracking protection approaches are discussed by Bender et al. (2016). We follow them with regard to the identification of selective prevention approaches to be most suitable for combining systematic tracking prevention with good user experience. The selective prevention approach follows the identify and block idea. All external referenced images need to be classified as either being tracking or non-tracking images. While non-tracking images remain, tracking images are either deleted or exchanged to preserve the e-mail format.

11.4.1 High-level design

To provide the functionality of the selective prevention approach, the combination of tracking detection and e-mail modification is required. Providing this functionality can be realized on the client-side as well as server based. Table 11.1 gives an overview of criteria that distinguishes the two approaches to decide for the more suitable one with regard to the solution objectives set.

While server- and client-based approaches share characteristics, many differences exist that are to be considered when designing a tracking prevention software solution. The server-based solution is installed on the mail server and therefore tightly integrated with the central point of mail reception. Since the relevant functionality in bundled on the server-side, the solution is client independent. Modifications and specialized functionality are not required on the

Table 11.1: Comparison of client-and server-based approaches

Evaluation Criteria	Server-based solutions	Client-side solutions
Sufficient solution available	No	No
Exhausting protection	Yes	If supported and active on all clients
Necessary setups	Once	For each device / mail client
Multi-platform support	Not required	Required
User configuration effort	None	On each device
Profits from heavy traffic	Combining information gives detection enhancements	Heavy load on client device

front-end (client) side. On the contrary, the client-side solution requires each device and mail client to realize the detection and modification procedures. If at least one mail client does not realize this, protection is insufficient. Considering the many different mail clients, platforms, and devices, the client-based approach requires extensive development efforts compared to the server-based solution. The availability is similar for both solution types. Former studies identified client-side solutions as rare and insufficient (Bender et al., 2016). Considering the quality of protection, the server-based approach can profit from receiving similar mails. Tracking mechanisms can be identified by analyzing differences between similar mails as tracking approaches require uniqueness identification of e-mail recipients (Fabian et al., 2015). To sum up, the server-based approach fulfills the requirements and demands to a greater extent, especially in the professional context, as enterprise-infrastructures involve centrally managed mail infrastructures. We therefore focus on a server-based solution in the realization.

11.4.2 Software specification

For ensuring the high-quality demands of such complex software, a well-founded engineering process was applied to operationalize the design and development phase of the design science research process. A model that matched the nature of the task is the *Rapid Application Development* Model by J. Martin (1991). Its development lifecycle is designed to enable faster development and higher quality results than traditional process models (Despa, 2014).

We design our framework as a flexible *micro-service* architecture. The process and data flow are shown in Figure 11.2. A detailed description is provided in the subsequent sections.

A system consisting of separate services brings about additional challenges for communication and security. Use of the Docker platform (Cook, 2017) and its container-based system support the objectives of universality, user-friendliness, flexibility, and plugin architecture on a system level. Docker also provides a local network that can be used for communication between running containers (Figure 11.3).

An advantage of this software design is that the services are only loosely coupled through a communication protocol and can be easily replaced. Further advantages of the decomposition

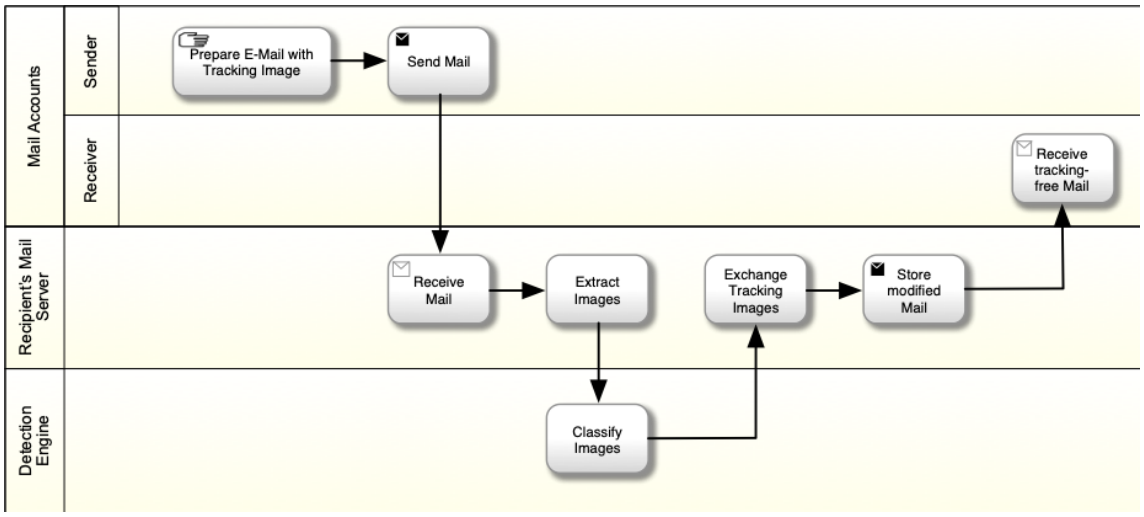


Figure 11.2: Process design and data flow in the software framework

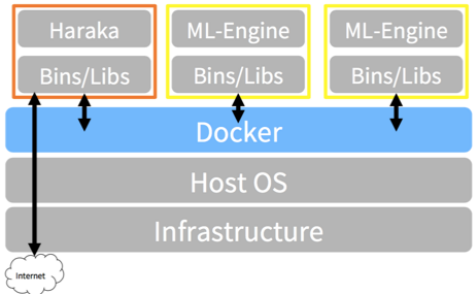


Figure 11.3: Detailed architecture with load-balancing of the detection engine

of an application into services are high system modularity, treatment of services as black-boxes, which makes the application easier to understand, and allows parallelized development, which is of special importance for large applications with multiple teams.

11.4.3 Technological Building Blocks

Node.js is a cross-platform JavaScript (JS) runtime-environment that can execute JavaScript code on the server. It was chosen for this project due to multiple reasons. First, it comes along with an active and big ecosystem that provides a lot of libraries and frameworks for uncommon issues. Second, it prevents a possible bottleneck (connecting to the detection engine) through its asynchronous I/O ability. Third, it is lightweight and runs on low resources, but is still very performant and scalable at the same time. Fourth, it considers possible open source contributions, since JavaScript and Node.js are widely known technologies with a large developer pool. Based on the project's objectives and requirements, one can see that especially the technical requirements like scalability, performance, and low resource usage could be met by utilizing Node.js. Version 8.6.0 was used for development, but compatibility with updates for the long-term supported version 8 is ensured.

Haraka is an open source Simple Mail Transfer Protocol (SMTP) Server written in JavaScript using the Node.js platform. It is built as plugin architecture around a very lightweight SMTP

core, which provides software engineers to hook into the mail processing. This flexible architecture paired with the event-driven access on the SMTP processing, grants developers all options to shape the server's behavior for their needs. Haraka is selected for this software framework due to ideally matching the requirements. It has high performance, allows customized behavior, has a plugin architecture from the base on, supports hook-ins into the SMTP processing, developer friendly, comes along with security features, and foremost is it not bound to any particular e-mail service provider.

R is a widely established open source framework for statistical calculations and machine learning. Together with Python, R can be seen as a standard platform for contemporary data science. While some studies criticize R for bad memory management (Krill, 2015), a completely rewritten version of the R core was recently provided by Microsoft[1] and overcomes corresponding concerns. However, we acknowledge that several interesting alternatives for implementing the machine-learning-based detection engine exist and continue to emerge in the form of new programming languages for data science applications (e.g., Julia) or user-friendly graphical machine-learning platforms (e.g., DeepLearningStudio[2]), which generate deployment code in some language. With these considerations in mind, we realize the detection engine as a plugin, which communicates with the e-mail server via HTTP using the package jug. The plug-in design makes it easy to replace the detection engine, which we currently implement in R, with some alternative technology upon need.

Docker is open source software that virtualizes an operating system (OS) for cloud applications, which are running in containers and are therefore isolated from the actual OS and from each other. It provides a lightweight layer of abstraction between the host OS and the containers, which enables certain functionalities such as setup automation and a separate local network for the containers. Docker utilizes resource isolation features of the Linux kernel to allow independent containers to be executed in a single Linux instance. This avoids the overhead of starting and maintaining a virtual machine (Vohra, 2017). Docker-compose is a feature of the Docker platform that allows defining and running multi-container applications. It performs the configuration, creation, and start-up process for all of the application's containers with a single command. The convenience of Docker leads to a high adoption in the software industry (Arijs, 2016; Vohra, 2017).

Docker was chosen for the framework due to the alignment of its features with the objectives and requirements. Especially its plugin-oriented architecture, flexibility, universality, user-friendliness, maintainability, extensibility, scalability, and security demands made Docker the platform of choice. The desired plugin architecture is supported through the container system as well as extensibility and scalability, flexibility and universality are given because Docker runs on every Linux-based OS and Windows. Security is enhanced since Docker-compose provides an internal network for containers to communicate. Only the container running the e-mail server is exposed to the Internet, while the containers with the detection engines are not. Therefore, researchers and developers not have to consider network security when creating new detection

[1] https://mran.microsoft.com/open
[2] https://deepcognition.ai/

engines. With a one command setup and start the whole application ecosystem and providing scaling out of the box, boosts user experience and adoption potential of the framework.

11.4.4 Data Flow

Figure 11.1 shows the top-level data flow in the framework. It starts with the sender sending an e-mail, which is routed via an e-mail server that implements our software. The Haraka SMTP server registers the incoming mail and executes authentication checks; if they pass, it starts receiving the e-mail. After all data is received, an event is emitted and the customized tracking prevention plugin "hooks" into the processing of the e-mail. As a first step, the body and the headers of the e-mail are parsed and handed over to the e-mail extractor function. This obtains the images with all their attributes from the large HTML-string, and prepares the headers to be passed along with the images to the detection engine.

Then, the mail extractor derives additional data and returns an array consisting of all images of the e-mail. The returned array is handed over to the communication module that sends the image objects to the detection engine. The mail server is not blocked while the detection engine processes the images. Consequently, the plugin can process the next e-mail until the response from the detection engine arrives.

When the answer containing all tracking images arrives, their source needs to be replaced in the e-mail body. Since the body of the e-mail is a large HTML-string, regular expressions (regex) can be applied. If a matching string is found, it is replaced by a new image source. Finally, the e-mail body is replaced with the tracking-free version and the e-mail is forwarded to its recipient.

Executing all of these steps without intermediary checks could be wasting computing power and makes the system more error prone. For example, if an e-mail does not contain any images, it should not run through the whole process, but rather be directly passed to the mail forwarding function. The same holds if the detection engine did not find any tracking images. Figure 11.4 displays the corresponding control flow of our software as UML activity diagram.

Analyzing a software system based on time, communication and execution provides an important perspective for a deeper understanding. In Figure 11.5, the sequential execution and communication of the framework is shown using UML. Stick arrowheads represent asynchronous messages, and the triangle head stand for synchronous messages.

In the diagram, the process of the framework is started when an e-mail arrives. As a first step, the mail extractor module is called with the e-mail body and headers as arguments. After the function returns, it is checked whether the e-mail contained images, which are then passed via the communication module to the detection engine. When the response from the detection engine arrives, it is checked if tracking images were detected. If so, the tracking images' sources are replaced with the link to a warning image. After completing the replacement, the e-mail is forwarded to the actual recipient inbox.

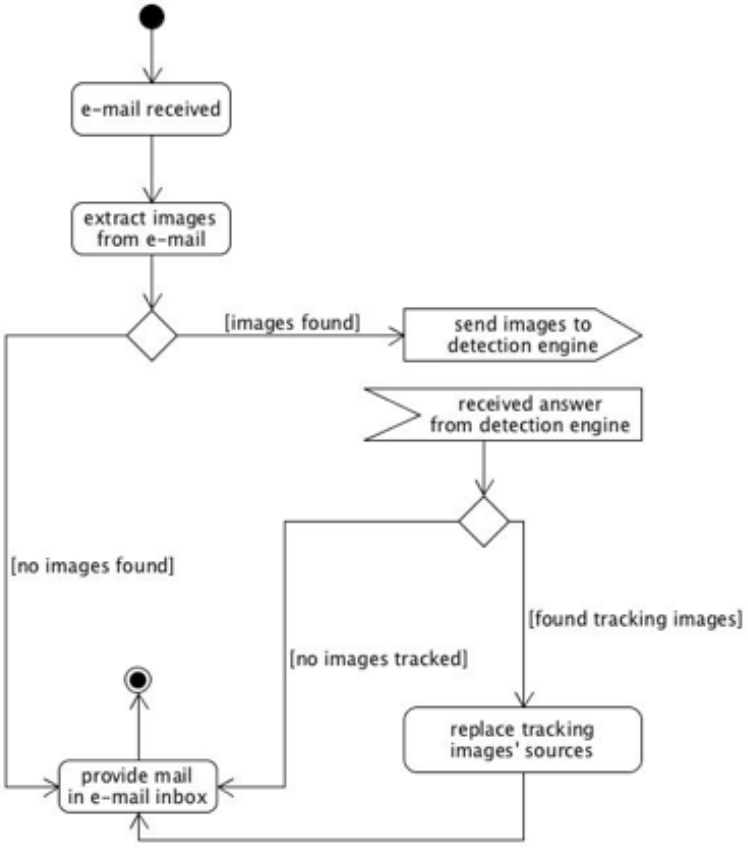


Figure 11.4: UML activity diagram of the software framework

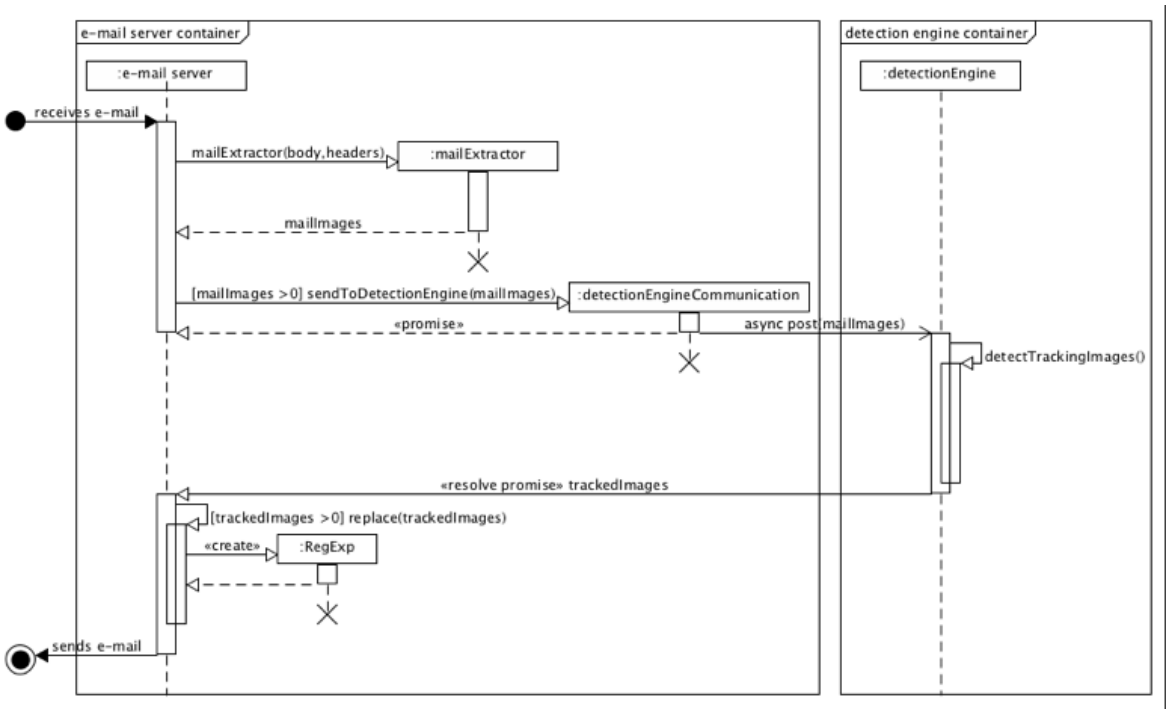


Figure 11.5: UML sequence diagram of the software framework

Table 11.2: Performance of classification models for tracking image detection for model-specific decision thresholds optimized on the training data

	Sensitivity	Specificity	AUC
Blacklist	0.29	0.98	0.637
Logistic regression	0.98	0.72	0.972
Random forest	0.92	0.98	0.994

11.4.5 Detection Engine

We build the detection engine on a machine learning classifier that identifies individual tracking images based on input features extracted from the email code (Haupt et al., 2018). The features are calculated from the HTML code, since loading the images is prohibited by the application, and fall into two groups. First, we capture the formatting of the image reference URL, since tracking image references typically show patterns of being generated and managed automatically. Indicators of these patterns are, for example, the file format or calculating statistics on the length, number of folders and changes between letters and numbers. Second, we relate each image to the other images within the same email, since separation of tracking and content management infrastructure manifests itself in distinct patterns within the URL. Tracking image links are accordingly served from different domains or have different folder structures or file formats than, for example, advertising content within the same email. All features are selected to be resilient against active manipulation by the tracker, which disqualifies absolute image size and keyword matching used in previous studies.

The detection engine is set up as a combination of the feature extraction module and any state-of-the-art classifier module to allow convenient updating of features and the classifier over time. For this study, we evaluated a logistic regression and a random forest model (Hastie et al., 2009) on classification accuracy and execution time to identify the binary classifier that is best suited to classify images as "tracking" or "non-tracking" based on the proposed features. The available data consists of newsletter emails collected from 300 companies in a 20-month period from 2015 to 2017 in a controlled experiment and contains 794,519 external images within 23,602 unique emails. We tune the detection models using five-fold cross validation on a training set of emails from a 5-month period in 2015. To ensure robust performance of the classifiers out of sample, we evaluate the classifiers on 30 randomly selected companies, whose emails were excluded from the training data. To further ensure robust performance over time, we include only emails received after the training data in a 15-month period from Nov 2015 to Jan 2017. We repeat the sampling and model training process ten times and report the average results over all company samples.

The proposed selective prevention system is intended to block only specific tracking images rather than all images in an email in order to inhibit the user experience as little as possible, while ensuring a maximum level of user privacy. However, the implicit cost of failing to block a tracking image and allowing a breach in user privacy is higher than falsely blocking content

images and impeding readability. We tune the probability threshold (for each classifier individually) on the training data set to account for the cost imbalance and assess the performance of the classifier in terms of the sensitivity and specificity, i.e. the percentage of tracking and non-tracking images that are correctly classified, respectively. For the empirical evaluation, we identify the probability threshold to be the value that maximizes the specificity of a classifier at a fixed sensitivity of at least 99.99% on the training data, based on the believe that many users have a strong preference for privacy. Future implementations could consider the sensitivity/specificity tradeoff as a user choice and set the threshold accordingly. Having fixed the sensitivity of each classifier on the training data, we compare detection engines w.r.t the mean sensitivity and specificity over the ten test sets described above and report the area-under-the-ROC-curve (AUC) for completeness (Table 11.2).

We observe that both statistical learning models outperform a blacklist approach based on known tracking providers and conclude that logistic regression and random forest are effective at detecting tracking images in a real-world setting. Regarding user experience, the logistic regression correctly classifies 76% of non-tracking images, while the random forest classifier correctly identifies over 99% of non-tracking images and thus leaving the email content completely intact for most emails. We adopt the random forest in our current implementation of the detection engine and stress its comparability to other state-of-the-art machine learning classifiers in terms of model complexity and prediction delay.

11.4.6 Scalability

Using Node.js as the underlying technology of the e-mail server provides good scalability. However, Node.js is a single threaded technology, which imposes certain restrictions. The Node.js community found a remedy in building a native cluster solution, where an orchestrating master process is spawned with the potential of starting as many workers as the system has central processing units (CPU). Haraka inherits this cluster technology.

The detection engine is a service that should be scaled when experiencing heavy load. To ensure proper usage of resources, a load balancer is placed before the detection engine containers. The load balancer distributes incoming messages from the e-mail server according to the free capacity of the detection engines. This mechanism keeps response rates of the complete system to a minimum and allows it to even handle huge amounts of traffic. If required, also the e-mail server can be multiplied. Recommended is to also place a load balancer in front of the e-mail server containers when scaling. Another feature of the system is that any scaling can be executed on startup, or even when the system is running, with one simple command: `docker-compose up --scale detectionengine=2`. This command starts the system with two detection engine containers as shown in Figure 11.2.

In summary, the system provides all required functionalities to serve on a large scale. Due to this fact, enterprise-grade usage and ease of use are supported.

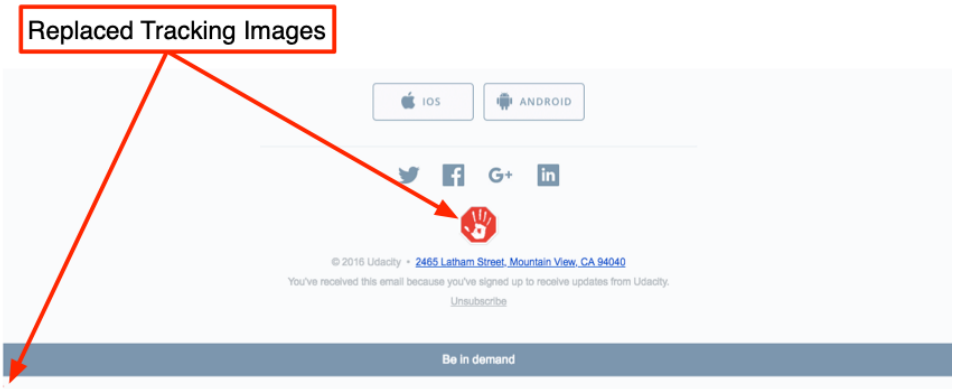


Figure 11.6: Tracking images replaced in example e-mail

11.5 Demonstration & Evaluation

11.5.1 End-User Perspective

For demonstration, an excerpt of a real-world e-mail is shown in Figure 11.6, an HTML newsletter from an online education platform. Its content contained tracking images that were filtered by our application.

One larger tracking image (a logo) as well as a tiny tracking pixel (at the lower left end of the e-mail) are detected. The red warning sign in Figure 11.6 demonstrates how a tracking image is replaced from an end user's perspective. It also illustrates that styling of mail content remains intact. Using this technique retains user experience in contrast to the intrusive approach of preventing all image downloads.

11.5.2 Software Complexity

An analysis of the software's complexity clarifies how efficiently an algorithm or piece of code operates, which is important in practice when system operations are time or resource critical. Accordingly, the algorithm's CPU (time) usage, memory usage, disk usage, and network usage should be taken into account. Disk and network complexities are negligible for the framework because network traffic is limited to sending from the mail server to the detection engine and back and no disk usage is implemented.

The e-mail server executes the image extraction, sending to the detection engine, and replacing of tracking images' sources. Sending the images is constant in its time complexity and linear in space complexity because the memory required depends on the amount of images in the mail body, but it has no influence on the amount of messages to be send. The mail extractor is a more complex piece of software. It needs to iterate over the HTML-string to extract the images and iterate over the image objects multiple times to destruct styling, calculate similarities etc. Clearly, a decision had to be made to either keep space complexity low at the expense of time complexity or vice versa. Due to e-mail not being a time-critical service, the first option was selected so quadratic time complexity was introduced in order to achieve a logarithmic space complexity. An advantage is that the objective of running on low resources can be met even in

high load situations. Delay would then increase, but the used memory will stay more or less the same.

In pre-studies, we established that the detection engine is able to classify images with high speed, though there is room for performance improvements, as the experimental section will show. More time is needed with the training of the machine learning models. However, this can be conducted offline and new models can be rolled out to the detection engine in a production environment at regular intervals.

The last operation is the replacement of tracked links in the images' sources. Through utilizing a regular expression, a linear time complexity with a constant space complexity is achieved. Linear time complexity results from iterating over the HTML-string just once and constant space because the pattern as well as the replacement is stored as string. Other options would have introduced a lot of complexity to the code and probably achieved a worse result.

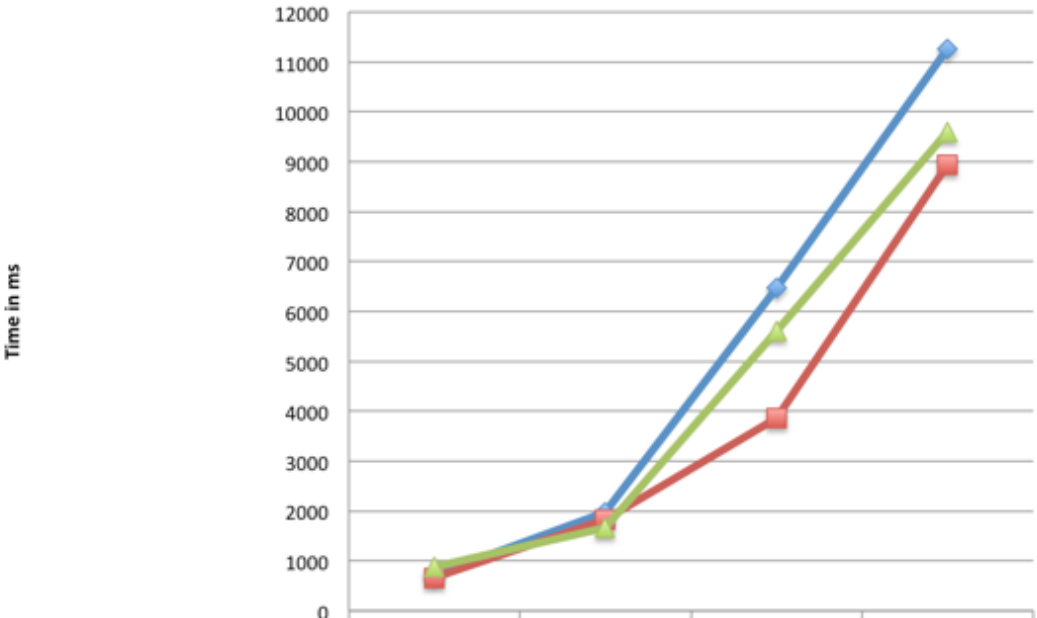
Overall, the complexity assessment also shows that the application is more sensitive to a large number of images in the e-mail body than to heavy traffic. But if the system is under heavy load with mails containing loads of images, the system will just slow down due to time complexity issues but will not break because of exceeding memory space.

11.5.3 Performance Experiments

Conducting performance experiments is an essential measure of the framework's suitability for production. For application in an enterprise environment, the responsible engineer needs to know how many instances and resources are needed to handle the traffic of the company.

The tests are operated using the Apache JMeter tool on an Apple MacBook Air Mid 2012 with a 1.8 GHz Intel Core i5 (i5-3427U) and 8GB 1600 MHz DDR3L RAM. Setup and running the software framework was conducted using Docker. Testing on a laptop and not a dedicated production machine serves as performance baseline. Focus is put on the analysis of multiple test settings to simulate real traffic and to inspect performance behavior for different variables. Scaling scenarios are realized through putting heavy load with a real traffic simulation on the system.

First, the response times of the system are analyzed for the combination of feature extraction plus a random forest classifier. Different e-mail contents are mixed to monitor behavior towards the number of images since the complexity analysis in the previous section indicated sensitivity towards this variable. Three tests were developed, real traffic simulation, meaning a mixture of one, twenty, and no images in the mail content. Secondly, one and no image mails mixed and finally, twenty and no images mixed. All tests were executed with a different amount of concurrent connections ranging from 2 to 40. Thereby, the test tool opens a new connection whenever another one has closed in order to keep a constant amount of connections open. For the first test plan, the framework was running on one container per service: one mail server and one detection engine.



	conc 2|3	conc 6	conc 20	conc 40
20 imgs, mixed with no images	712	1962	6472	11267
1 img, mixed with no images	664	1829	3863	8944
all mixed, real traffic sim	874	1670	5608	9599

Figure 11.7: Average response times on a single instance with mixed traffic

The results show a nearly linear relationship between concurrent connections and average response time in all tests (Figure 11.7). Comparing numbers, the average response time approximately doubles if the amount of concurrent connections doubles.

This is a very promising result and indicates that the system has not reached its limits yet, since no exponential growth of response time was observable. Being able to handle 40 concurrent SMTP-connections with ten seconds delay on average already proves the framework to be suitable for midsize companies with this single instance setup.

Furthermore, it has to be mentioned that during all tests the system had an around 1.5 e-mail throughput rate per second. Speed alone is not the only important indicator for the framework; also the error rate has to be considered. As previously mentioned, Node.js and Haraka become slower rather than throwing errors. This is supported through the tests by achieving a 0.00% error rate.

In the previous section, it was estimated that the number of images in the e-mail body impacts delivery delay due to a quadratic time complexity, which is confirmed by Figure 11.7. It shows that the test with just one image has the lowest average response time, while the test with twenty images resulted in the highest. Real traffic simulation is situated in between, which meets the expectation since the test sent twenty, one, and no image-containing e-mails. Ultimately, it is clear that the framework's performance is strongly influenced by the number of images in the e-mails it is processing.

Monitoring system resources is important to check if the low resource objective is fulfilled by

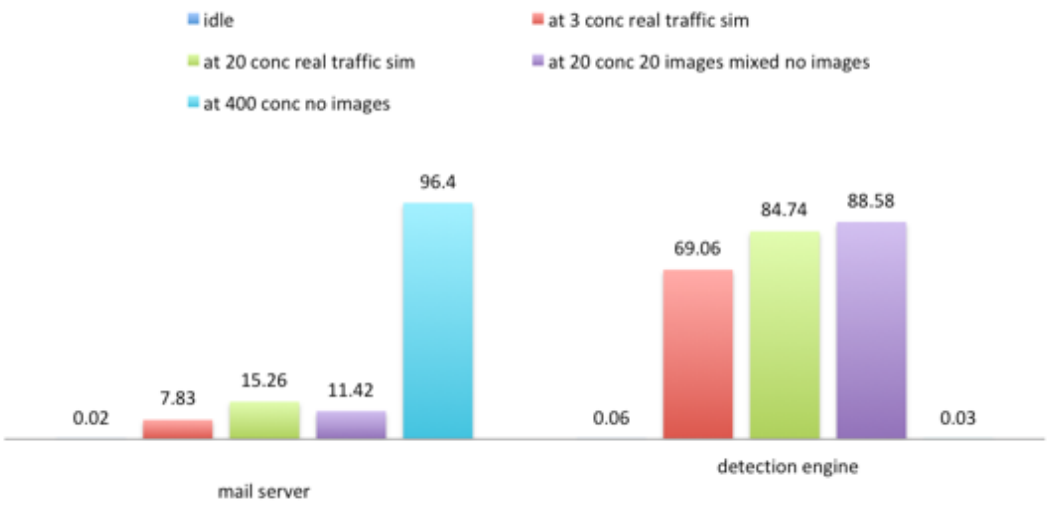


Figure 11.8: CPU usage in percent during experiments

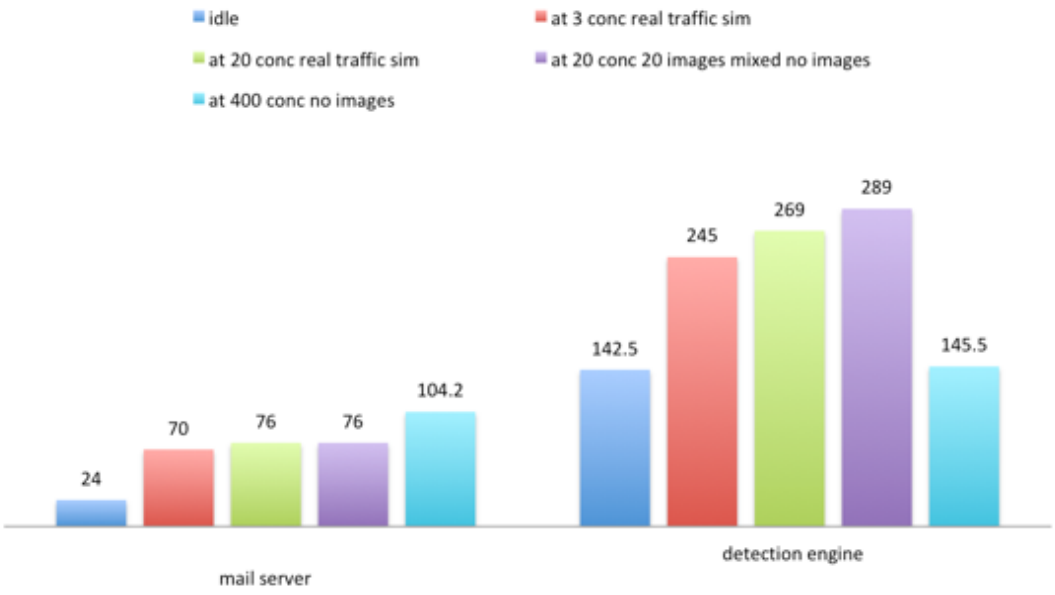
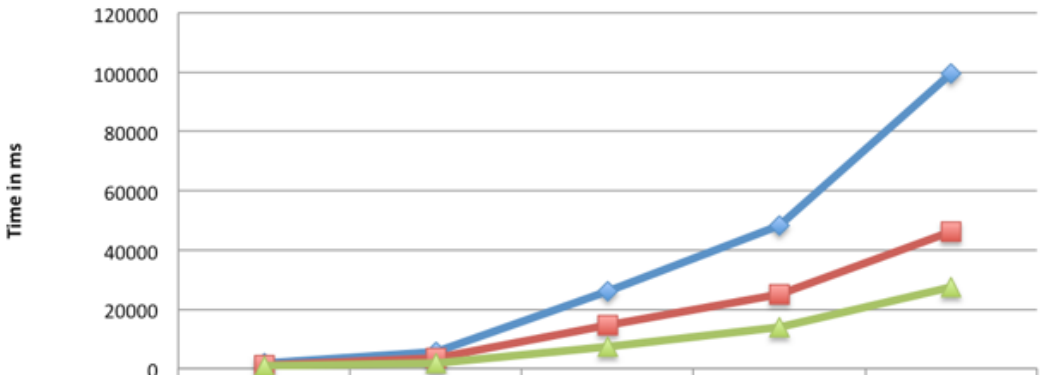


Figure 11.9: Memory usage in megabytes during experiments

the system. During test execution the system metrics were measured and resulted in the data shown in Figures 11.8 and 11.9. Figure 11.8 clearly reveals the CPU distribution between server and detection engine. Depending on the task, the usage share is flexibly distributed; when more cycles are needed for the detection engine the mail server lowers its share. It is observable that the CPU is not fully claimed on three concurrent connections. An interesting difference exists between 20 concurrent connections with real traffic and 20 concurrent connections with a higher image load. Recalling the faster response times of the real traffic simulation from Figure 11.7, this can be explained through allocation of more CPU resources to the mail server, whereas the detection engine required more CPU power on higher image load, leading to slower response times.

Figure 11.9 supports this argument; the detection engine required more memory during high



	conc 6	conc 20	conc 120	conc 240	conc 480
Non scaled	1670	5608	26128	48261	99442
scaled DE=2	1049	3560	14629	25055	46143
scaled DE=4	921	1916	7547	13933	27400

Figure 11.10: Average response times on scaling, simulating real traffic

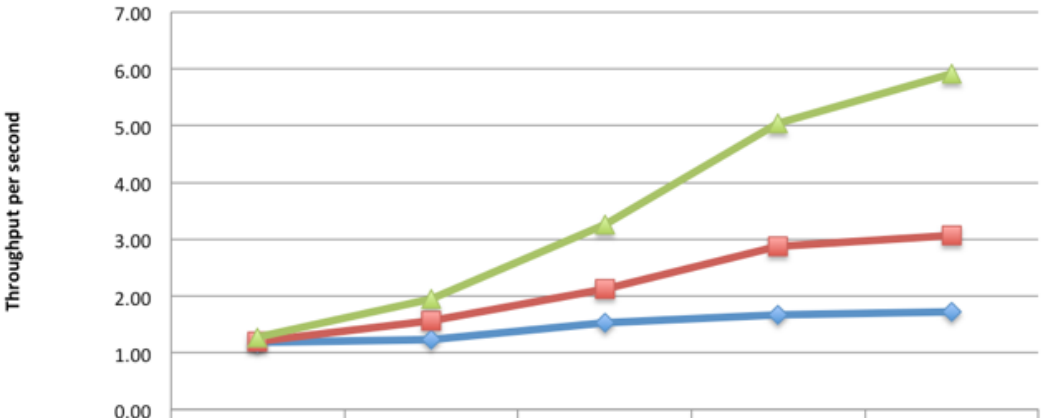
image loads and keeping the mail server's memory stable at 76 MB. Accordingly, the image sensitivity issue is backed through the system monitoring metrics.

In addition to the image traffic tests, a standard load test was performed to measure the general behavior of the framework when handling e-mails without images. The measured system metrics support the complexity analysis. When exposing the framework to the enormous amount of 400 concurrent connections, it reacts by using a lot of CPU cycles, but stays low on memory resources. So, the framework behaves exactly how the complexity analysis predicted. In this heavy load test the software artifact achieved a 9.8 seconds average response time, 0.00% errors, and an impressive throughput of 35.76 e-mails per second. These numbers show the performance of the framework when not having to handle image detection. One has to remember that these numbers were achieved running just one e-mail server- and one detection engine container.

When switching from one instance per service to multiple service containers, at first the bottleneck service needs to be identified. Looking at the previous results, it is straightforward to conclude that the framework can be just as fast as the detection engine. So, there is one issue in service-oriented systems: the whole is just as fast as its slowest part. Having loosely coupled services allows simply multiplying service instances to achieve horizontal scaling. This mechanic mitigates problems of bottleneck services. As a consequence, the detection engine was scaled up to two and four instances, and the test load was further increased through more concurrent connections.

Analysis of Figure 11.10 reveals that the framework's response times are growing as a polynomial, but the effect is moderated for the scaled instances. Linear behavior as in Figure 11.7 can be also observed here. Response times are linearly correlated with the number of detection engine instances and therefore increase with a smaller ratio. This behavior was expected.

It is surprising that the single detection engine system manages to handle 480 concurrent connections with growing just to an average response time of one minute and forty seconds,



	conc 6	conc 20	conc 120	conc 240	conc 480
Non scaled	1.18	1.23	1.53	1.67	1.72
scaled DE=2	1.20	1.57	2.13	2.88	3.08
scaled DE=4	1.28	1.95	3.26	5.04	5.92

Figure 11.11: E-mail throughput per second, simulating real traffic

which is yet an acceptable speed for asynchronous e-mail. Again, there were no errors returned for any request in the tests and even the result of the single instance framework was far away from typical sender SMTP timeouts that range between two to ten minutes (Braden, 1989, Section 5.3.2). Multiple detection engine instances provide no advantage for few concurrent connections, especially when put into context with the increased resource usage.

On the other hand, the horizontal scaling shows its true potential when having heavy concurrent traffic. For example, the system with four detection engines at 480 concurrent requests takes as long as the single system takes for 120 or the doubled system for 240. The slightly higher response time average of the scaled framework with four detection engine instances probably originates from the e-mail server stressing to handle 480 connections at a time. Delay caused by the load balancer is negligible and thus can be excluded as a cause.

Looking at Figure 11.11, it is directly recognizable that scaling increases the e-mail throughput. Again, a linear relationship between the results for each system is recognizable. However, the shapes of the lines reveal an increasing slope from 20 to 240 concurrent connections, but a decrease in growth rate from 240 onwards. Having eliminated the detection engine bottleneck through scaling, this implies that the e-mail server is approaching its limits and has already a restricting impact at 480 concurrent mails. Problems with reaching the e-mail server's limits can be easily alleviated through placing another load balancer in front of it and multiplying mail server instances.

As an extreme use case, a company might handle 20 million e-mails a day, which are 230 mails per second (Sergeant, 2017). A single Haraka instance can handle more than 35 mails per second without extraction; with extracting images, 17 mails per second were measured. To provide a buffer, assume a throughput of 15 mails per second. Then it would require just 16 e-mail server containers to handle the full traffic of that company. Assuming further linear behavior, it requires 12 detection engine instances per Haraka instance. In total it would require 16 e-mail

server instances with 196 detection engine instances to filter and route the complete e-mail traffic of the company at reasonable speed.

Slightly over-scaling provides enough buffers for peak traffic, but still the service would just need more time but not throw errors when being overloaded. Having a 1:12 ratio of mail server to detection engines recommends future improvement of the detection engine's performance. Also, the 16 Haraka instances might be scaled down through separating the mail extractor from the server to a dedicated service.

Regarding the replacement of links, the regular expression method is commonly seen as one of the fastest methods. However, it can be discussed if the replacing should be executed on the server or rather also be handled via its own micro-service. These two measures could cut the number of required mail server instances by half.

Evolving to a full micro-service system would be beneficial for scalability, but would also add complexity on a service level. Then, HTTP may be replaced by using web sockets since communication frequency would increase and this switch of protocols could reduce overhead. This approximation shows that the framework is prepared for extremely large-scale enterprise usage, although the detection engine would require improvements to reduce the required number of instances.

Finally, it can be concluded that all defined objectives and technical requirements are fulfilled. Especially the limits of maximum five minutes e-mail delay, handling more than 20 concurrent connections with less than 250 MB memory usage, was exceedingly satisfied.

11.5.4 Static Code Quality

Code quality is a metric that does not affect the execution of the program in any way, but it does affect the further development of it greatly. However, code quality is always hard to assess and depends on the individual style of the software engineer. Nonetheless conventions, best practices, and style guides were developed in an attempt to streamline software development. Tools such as linters and integrated development environments (IDE) support developers to apply them during implementation. Most of these tools also offer analysis of files or directories. They point at errors, emit warnings, and make suggestions. Developers are not obliged to integrate suggestions and warnings.

Code quality analysis was executed through using the JetBrains *Webstorm* JavaScript IDE and the *ESLint* plugin. Results of the static code analysis indicate a high code quality level and application of conventions and best practices as well as utilization of a proper code style. This high standard is also expected by the open source community in order to attract contributors for further development.

11.6 Discussion

Facing that 92% of all e-mail openings of commercial newsletters might be unprotected against tracking (Bender et al., 2016) shows the scope of the privacy problem. Concerning the differentiation to other existing protection services against e-mail tracking, there are a few other services that broadly address the same issue, but always with problems; either it was dependency on a specific browser or a very complicated procedure to setup the service. Summarizing, earlier privacy solutions either lack user friendliness, precision or independency.

In general, it should be questioned why the user should be responsible for making his own e-mail inbox tracking free. Why is not the service provider offering better privacy protection? Furthermore, companies might want to protect their employees' work e-mail addresses from tracking. A simple browser or inbox plugin could also be attractive to end-users, but it would always come along with a dependency on a third-party application. Our goal was to remain independent and let many users benefit if a mail service implements our framework.

The motivation of this study was to create an enterprise-grade framework that fills that gap in the field of privacy-enhancing technologies. Recalling the evaluation section, it can be clearly said that the framework fulfills these criteria. The framework is ready and suitable to be applied in actual production conditions. Metrics from the evaluation sections show that this is achieved through right choices regarding technologies and a solid software design. The ability to handle more than 400 concurrent connections with reasonable response times on low resources and with a 0% error rate proves the framework's quality.

However, meeting an end-user demand is not the only purpose of this software. It was also built to provide a basis for further research in the field of e-mail tracking. The platform aims to be used by researchers to test new detection engines, gather data about the problem itself, and to serve as a real-world test environment. Additionally, the selective filtering approach was demonstrated as valid through implementation in a real application. This framework augments the research field by a practical and extensible system to support investigations of any direction.

Concerning limitations, there are aspects that would also have supported the decision in favor of monolithic application architecture. However, monolithic designs are known for their struggles with scaling. Although the architecture would provide an environment for easier logging and features such as shared memory, these advantages do not outweigh the benefits of a micro-service architecture. As tracking being a rapid evolving domain, the flexibility of micro-services allows exchanging parts without the need to redesign the whole solution, which fosters further development.

A similar, but contrary argument could go even further in the direction of micro-services: Why is not every task decoupled from the mail server? This seems like a compelling avenue for future work; the decoupling of the mail extractor and the replacing would free resources on the mail server. On the other hand, this would lead to higher communication traffic in the local network and thus would the HTTP overhead be higher. For sure, this can be easily solved with

a switch to web-sockets. Another point against it is that there are fixed resources required by every container such as a runtime environment and libraries as well as a new load balancer if multiplication is planned. Trade-offs had to be made and due to the framework prototype stage, it was chosen to keep the extracting and replacing task at the server. For upcoming refactoring iterations, it is recommended decouple both operations into separate services.

For a detection engine that has to handle heavy traffic, needs to process HTTP, and should run on limited resources, R is a suboptimal choice. Consequently, future detection engine development should consider execution speed, complexities, and resource usage. In the field of machine learning there are alternatives such as Python or Julia, which are also popular and proven in large-scale applications.

Yet another aspect is the discussion if JavaScript should be used in such specific use cases as e-mail. The community is divided in that point. If technologies emerge that bring more advantages than Node.js to the table, then a mail server re-implementation could be recommended. For now, it was the most efficient solution to develop the prototypical artifact in the timeframe and scope of this article.

Concerning future research, next logical steps for the framework are numerous and should be tackled in separate studies. First of all, the detection engine should be improved in order to even better meet the identified requirements of the framework. It would be interesting to know if there is a more efficient method than the currently used one or even a possibility to split the detection engine into different tasks with certain checkpoints. Hand in hand with a rework, but probably as a separate study, would be the extension of the detection engine and mail extractor to not only check for tracking images, but tracking links in general. Often, users are clicking on links by accident and lose all their tracking protection through this mistake.

A further next step would be the enterprise-grade application of the framework. Gathering data about the framework's behavior in a non-experimental setting would contribute to future improvements. Real usage data has also a positive effect on adoption by the open source community.

An extension that would foster future privacy research would be the implementation of a proper logging engine and or a sophisticated analytics tool. These two features would be of high value for quantitative research in e-mail tracking. Metrics such as the tracking images' domains and other characteristics would become analyzable. Also, the resulting database from these two services would be beneficial to the general e-mail research community.

To foster application and future research, we will publish the source code of our framework in an open repository such as *GitHub* under an open source license. Together with this article, this will support *communication*, the final step in the formal process of design science research (Peffers et al., 2007).

11.7 Conclusion

Tracking methods, such as web and e-mail tracking, are popular marketing tools. Data derived from tracking provides valuable information regarding a person's interests and reception behavior. Modern e-mail tracking methods allow the sender to determine how often an e-mail was opened, the device used to read the e-mail, which links were clicked, and the location and time when the recipient opened an e-mail. While different aspects concerning e-mail tracking have been studied in isolation, a study integrating former efforts in a ready to use countermeasure is still missing. This contribution addresses this gap, by conceptualizing and implementing a novel protection framework against e-mail tracking.

Following the Design Science Research Method, we develop a software being capable to identify tracking images in e-mails via machine learning with very high accuracy and can selectively replace them so that an untracked e-mail is provided for the end user without any manual effort. Our mail protection framework is developed as enterprise-grade software, flexibly extensible, highly scalable, and ready to be applied in actual production conditions. The experimental evaluation section shows that this is achieved through corresponding choices regarding technologies and the creation of a solid and flexible software design.

Bibliography

Alsaid, A., & Martin, D. (2003). Detecting Web Bugs with Bugnosis: Privacy Advocacy through Education (R. Dingledine & P. Syverson, Eds.). In R. Dingledine & P. Syverson (Eds.), *Privacy Enhancing Technologies*, Berlin, Heidelberg, Springer. https://doi.org/10.1007/3-540-36467-6_2

Arijs, P. (2016). Docker Usage Statistics: Increased Adoption by Enterprises and for Production Use.

Bender, B., Fabian, B., Haupt, J., Lessmann, S., Neumann, T., & Thim, C. (2018). Track and Treat - Usage of E-Mail Tracking for Newsletter Individualization, In *Proceedings of the 26th European Conference on Information Systems (ECIS'18)*, AIS.

Bender, B., Fabian, B., Lessmann, S., & Haupt, J. (2016). E-Mail Tracking: Status Quo and Novel Countermeasures, In *Proceedings of the 37th International Conference on Information Systems (ICIS)*, AIS.

Bhattacharyya, M., Hershkop, S., & Eskin, E. (2002). MET: An Experimental System for Malicious Email Tracking, In *Proceedings of the 2002 Workshop on New Security Paradigms*, Virginia Beach, Virginia, ACM. https://doi.org/10.1145/844102.844104

Bilos, A., Turkalj, D., & Kelic, I. (2016). Open-Rate Controlled Experiment in E-Mail Marketing Campaigns. *Trziste*, *28*(1), 93–109.

Bonfrer, A., & Drèze, X. (2009). Real-time evaluation of e-mail campaign performance. *Marketing Science*, *28*(2), 251–263. https://doi.org/10.1287/mksc.1080.0393

Bouguettaya, A., & Eltoweissy, M. (2003). Privacy on the web: Facts, challenges, and solutions. *IEEE Security & Privacy Magazine*, *1*(6), 40–49. https://doi.org/10.1109/MSECP.2003.1253567

Braden. (1989). *Requirements for Internet Hosts – Communication Layers* (RFC No. 1122). RFC Editor.

Bujlow, T., Carela-Espanol, V., Sole-Pareta, J., & Barlet-Ros, P. (2017). A survey on web tracking: Mechanisms, implications, and defenses. *Proceedings of the IEEE*, *105*(8), 1476–1510. https://doi.org/10.1109/JPROC.2016.2637878

Cook, J. (2017). The Docker Engine. In *Docker for Data Science* (pp. 71–79). Berkeley, CA, Apress. https://doi.org/10.1007/978-1-4842-3012-1_4

Despa, M. L. (2014). Comparative study on software development methodologies. *Database Systems Journal*, *5*(3), 37–56.

Englehardt, S., Han, J., & Narayanan, A. (2018). I Never Signed Up For This! Privacy Implications of Email Tracking, In *Proceedings on Privacy Enhancing Technologies*.

Ermakova, T., Fabian, B., Bender, B., & Klimek, K. (2018). Web Tracking: A Literature Review on the State of Research, In *51nd Hawaii Conference on System Sciences*. https://doi.org/10.24251/HICSS.2018.596

Ermakova, T., Hohensee, A., Orlamünde, I., & Fabian, B. (2017). Privacy-invading mechanisms in e-commerce: A case study on German tourism websites. *International Journal of Networking and Virtual Organisations*, *20*(2), 105–126. https://doi.org/10.1504/IJNVO.2019.097629

Fabian, B., Bender, B., & Weimann, L. (2015). E-Mail Tracking in Online Marketing: Methods, Detection, and Usage, In *12th International Conference on Wirtschaftsinformatik*, Osnabrück, Germany.

Farahat, A., & Shanahan, J. (2013). Econometric Analysis and Digital Marketing: How to Measure the Effectiveness of an Ad, In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, Rome, Italy, ACM. https://doi.org/10.1145/2433396.2433502

Fonseca, F., Pinto, R., & Meira, W. (2005). Increasing User's Privacy Control through Flexible Web Bug Detection, In *Third Latin American Web Congress*, Buenos Aires, Argentina, IEEE. https://doi.org/10.1109/LAWEB.2005.19

Goldfarb, A., & Tucker, C. (2012). Privacy and innovation. *Innovation Policy and the Economy*, *12*(1), 65–90. https://doi.org/10.1086/663156

Grimes, G. A., Hough, M. G., & Signorella, M. L. (2007). Email end users and spam: Relations of gender and age group to attitudes and actions. *Computers in Human Behavior*, *23*(1), 318–332. https://doi.org/10.1016/j.chb.2004.10.015

Gu, X., Yang, M., Shi, C., Ling, Z., & Luo, J. (2017). A novel attack to track users based on the behavior patterns: A Novel Attack to Track Users Based on the Behavior Patterns. *Concurrency and Computation: Practice and Experience*, *29*(6), e3891. https://doi.org/10.1002/cpe.3891

Hameed, S., Kloht, T., & Fu, X. (2013). Identity based email sender authentication for spam mitigation, In *Eighth International Conference on Digital Information Management*, Islamabad, Pakistan, IEEE. https://doi.org/10.1109/ICDIM.2013.6694015

Harding, W. T., Reed, A. J., & Gray, R. L. (2001). Cookies and web bugs: What they are and how they work together. *Information Systems Management*, *18*(3), 17–24.

Hartemo, M. (2016). Email marketing in the era of the empowered consumer. *Journal of Research in Interactive Marketing*, *10*(3), 212–230. https://doi.org/10.1108/JRIM-06-2015-0040

Hasouneh, A. B. I., & Alqeed, M. A. (2010). Measuring the effectiveness of e-mail direct marketing in building customer relationship. *International Journal of Marketing Studies*, *2*(1), 48–64. https://doi.org/10.5539/ijms.v2n1p48

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The Elements of Statistical Learning* (Second). New York, Springer.

Haupt, J., Bender, B., Fabian, B., & Lessmann, S. (2018). Robust identification of email tracking: A machine learning approach. *European Journal of Operational Research*, *271*(1), 341–356. https://doi.org/10.1016/j.ejor.2018.05.018

Herzberg, A. (2009). DNS-based email sender authentication mechanisms: A critical review. *Computers & Security*, *28*(8), 731–742. https://doi.org/10.1016/j.cose.2009.05.002

Javed, A. (2013). POSTER: A Footprint of Third-Party Tracking on Mobile Web, In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, Berlin, Germany, ACM. https://doi.org/10.1145/2508859.2512521

Jensen, C., Sarkar, C., Jensen, C., & Potts, C. (2007). Tracking Website Data-collection and Privacy Practices with the iWatch Web Crawler, In *Proceedings of the 3rd Symposium on Usable Privacy and Security*, Pittsburgh, PA, USA, ACM Press. https://doi.org/10.1145/1280680.1280686

Krill, P. (2015). Why R? The Pros and Cons of the R Language.

Leon, P., Ur, B., Shay, R., Wang, Y., Balebako, R., & Cranor, L. (2012). Why Johnny Can't Opt Out: A Usability Evaluation of Tools to Limit Online Behavioral Advertising, In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Austin, Texas, USA, ACM Press. https://doi.org/10.1145/2207676.2207759

Li, T.-C., Hang, H., Faloutsos, M., & Efstathopoulos, P. (2015). TrackAdvisor: Taking Back Browsing Privacy from Third-Party Trackers (J. Mirkovic & Y. Liu, Eds.). In J. Mirkovic & Y. Liu (Eds.), *Lecture Notes in Computer Science*, Cham, Springer International Publishing. https://doi.org/10.1007/978-3-319-15509-8_21

Luo, X., Nadanasabapathy, R., Zincir-Heywood, A. N., Gallant, K., & Peduruge, J. (2015). Predictive Analysis on Tracking Emails for Targeted Marketing (N. Japkowicz & S. Matwin, Eds.). In N. Japkowicz & S. Matwin (Eds.), *Discovery Science*, Cham, Springer International Publishing. https://doi.org/10.1007/978-3-319-24282-8_11

Martin, D., Wu, H., & Alsaid, A. (2003). Hidden surveillance by web sites: Web bugs in contemporary use. *Communications of the ACM*, *46*(12), 258. https://doi.org/10.1145/953460.953509

Martin, J. (1991). *Rapid Application Development*. New York, Maxwell Macmillan International.

Mittal, S. (2010). *User Privacy and the Evolution of Third-party Tracking Mechanisms on the World Wide Web* (Master Thesis). Stanford University.

Parra-Arnau, J. (2017). Pay-per-tracking: A collaborative masking model for web browsing. *Information Sciences*, *385-386*, 96–124. https://doi.org/10.1016/j.ins.2016.12.036

Peffers, K. E. N., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77.

Ridley-Siegert, T. (2016). DMA Insight: Consumer email tracking report 2015. *Journal of Direct, Data and Digital Marketing Practice*, *17*(3), 163–169. https://doi.org/10.1057/dddmp.2015.54

Sammons, J., & Cross, M. (2017). Email safety and security. In *The Basics of Cyber Safety* (pp. 75–86). Elsevier. https://doi.org/10.1016/B978-0-12-416650-9.00004-8

Sanchez-Rola, I., Ugarte-Pedrero, X., Santos, I., & Bringas, P. G. (2017). The web is watching you: A comprehensive review of web-tracking techniques and countermeasures. *Logic Journal of IGPL*, *25*(1), 18–29. https://doi.org/10.1093/jigpal/jzw041

Sergeant, M. (2017). Large Scale Haraka Users.

Sommerville, I., & Sawyer, P. (1997). Viewpoints: Principles, problems and a practical approach to requirements engineering. *Annals of Software Engineering*, *3*(1), 101–130. https://doi.org/10.1023/A:1018946223345

Stolfo, S. J., Hershkop, S., Wang, K., Nimeskern, O., & Hu, C.-W. (2003). A Behavior-Based Approach to Securing Email Systems (V. Gorodetsky, L. Popyack, & V. Skormin, Eds.). In V. Gorodetsky, L. Popyack, & V. Skormin (Eds.), *Computer Network Security*, Berlin, Heidelberg, Springer. https://doi.org/10.1007/978-3-540-45215-7_5

Tsalis, N., Mylonas, A., & Gritzalis, D. (2016). An Intensive Analysis of Security and Privacy Browser Add-Ons (C. Lambrinoudakis & A. Gabillon, Eds.). In C. Lambrinoudakis & A. Gabillon (Eds.), *Risks and Security of Internet and Systems*, Cham, Springer International Publishing. https://doi.org/10.1007/978-3-319-31811-0_16

Vaas, L., & Stockley, M. (2014). How Emails Can Be Used to Track Your Location and How to Stop It.

Vohra, D. (2017). Docker Services. In *Docker Management Design Patterns* (pp. 55–84). Berkeley, CA, Apress. https://doi.org/10.1007/978-1-4842-2973-6_4

Yamada, A., Hara, M., & Miyake, Y. (2011). Web tracking site detection based on temporal link analysis and automatic blacklist generation. *Journal of Information Processing*, *19*, 62–73. https://doi.org/10.2197/ipsjjip.19.62

Zhang, X. A., Kumar, V., & Cosguner, K. (2017). Dynamically managing a profitable email marketing program. *Journal of Marketing Research*. https://doi.org/10.1509/jmr.16.0210